Aus der Klinik und Poliklinik für Nuklearmedizin
der Universität Würzburg
Direktor: Professor Dr. med. Chr. Reiners

Evaluierung der Intra- und Interobserver-Variabilität bei der 2D-Ultraschall-Schilddrüsenvolumetrie an einem Schilddrüsenphantom – Vergleich zu 3D-Ultraschall-Referenzmessungen an gesunden Probanden

Inaugural-Dissertation
zur Erlangung der Doktorwürde der
Medizinischen Fakultät
der
Bayerischen Julius-Maximilians-Universität zu Würzburg

vorgelegt von
Paul Andermann
aus Amberg

Würzburg, März 2007

Referent:	Prof. Dr. Chr. Reiners
Korreferent:	Priv.-Doz. Dr. M. Beer
Dekan:	Prof. Dr. M. Frosch

Tag der mündlichen Prüfung: 05.07.2007

Der Promovend ist Arzt.

Der besten Familie,
die man sich wünschen kann,
in Liebe und Dankbarkeit gewidmet

INHALTSVERZEICHNIS

1. EINLEITUNG

1.1 Grundlegende Aspekte der Ultraschalldiagnostik

Ultraschall bezeichnet alle mechanischen Schallwellen mit Frequenzen oberhalb 20 kHz. In Natur und Technik wird er erfolgreich zur Ortung und Messung eingesetzt. In der Medizin untersucht man seit mehr als 50 Jahren Gewebestrukturen mit Hilfe des Ultraschalls. Die Qualität des Untersuchungsablaufs und der Befundinterpretation wird wesentlich bestimmt durch die Kenntnis und Anwendung physikalischer Gesetzmäßigkeiten. Sie bestimmen auch die Anforderungen an die verwendeten Geräte für die Untersuchung und Dokumentation. Schallvorgänge sind elastische Schwingungen von Materie. Im Gegensatz zur elektromagnetischen Welle ist die Schallwelle mit longitudinaler und transversaler Ausbreitung an Materie gebunden. Sie ist eine Folge von zeitlich und räumlich sich ausbreitenden Verdichtungen und Verdünnungen in Flüssigkeiten und festen Körpern. Diese Schwingungen können optisch sichtbar gemacht werden und stellen so die Basis für die Ultraschalldiagnostik dar. Mit ihr lassen sich die Echostruktur (diffuse, uni- oder multifokale Läsionen), die Echogenität (echonormal, echofrei, echoarm, echodicht) und benachbarte Strukturen sowie Raumforderungen im Halsbereich beurteilen. Heute werden für die Schilddrüsendiagnostik Ultraschallscanner mit einer Sendefrequenz von 5 bis 10 MHz empfohlen.

1.2 Klinischer Einsatz des Ultraschalls in der Schilddrüsendiagnostik

Sonographische Untersuchungstechniken sind in der Humanmedizin weit verbreitet und fest in der Primärdiagnostik etabliert. Aufgrund der guten Verfügbarkeit und des geringen Risikos stellt die Sonographie das wichtigste und weltweit am häufigsten verwendete bildgebende Verfahren in der Schilddrüsendiagnostik dar [Hegedüs 1998, Knudsen 1999]. Mehrere weitere Vorteile machen die Sonographie zur Methode der Wahl bei der morphologischen Bildgebung der Schilddrüse: sie besitzt allgemein eine gute Validität, ist nur wenig invasiv, leicht

und schnell durchzuführen und erlaubt auch eine Beurteilung der angrenzenden Halsweichteile; sie kommt ohne Strahlenexposition aus, da biologisch inert, hilft bei der Durchführung gezielter Biopsien und Feinnadelpunktionen und liefert viel präzisere und zuverlässigere Informationen über Schilddrüsenvolumen und –morphologie als die Palpation oder szintigraphische Verfahren, die nur eine grobe Volumenabschätzung erlauben [Langer 1989, Knudsen 1999, Hussy 2000, Van Isselt 2003]. Die Planimetrie aus der Szintigrammfläche birgt insbesondere Fehlermöglichkeiten aufgrund der zweidimensionalen Projektion der Schilddrüse ohne ihren Tiefendurchmesser, der fehlenden Darstellbarkeit nicht speichernden Gewebes und der unsicheren Definition der Randkontur [Igl 1981]. Daher spielt die Sonographie als objektive Methode für die Volumenbestimmung der Schilddrüse heute die zentrale Rolle. Außerdem ist sie kostengünstig, verglichen mit anderen bildgebenden Verfahren. In der Klinik für Nuklearmedizin der Universität Würzburg werden pro Jahr etwa 5000 Ultraschalluntersuchungen der Schilddrüse von ca. 10 Ärzten durchgeführt.

Hauptindikationen für den klinischen Einsatz des Ultraschalls sind Diagnostik und Therapiekontrolle benigner und maligner Schilddrüsenerkrankungen. So können bei einer Vielzahl von pathologischen Schilddrüsenbefunden wie z.B. Struma nodosa, M. Basedow oder Thyreoiditis morphologische Parenchymveränderungen diagnostisch wegweisend sein. Mit Hilfe der Sonographie ist eine Differenzierung zwischen diffusen Schilddrüsenerkrankungen und knotig veränderten Schilddrüsen möglich [Gallo 2003]. Bei Autoimmunerkrankungen der Schilddrüse sind im sonographischen Erscheinungsbild die abnormen hypoechogenen Veränderungen charakteristisch [Van Isselt 2003], während sich bei knotigen Herdbefunden eine große Variationsbreite zeigen kann (iso-, hyper- oder hypoechogen bis hin zu echokomplex). Nach einer aktuellen Metaanalyse bewegt sich die Prävalenz von knotigen Herdbefunden in der Schilddrüse, die zufällig durch Ultraschall entdeckt werden, zwischen 20% und 70% [Tan 1997].
Die Prävalenz von Schilddrüsenknoten in Deutschland ist weiterhin relativ hoch. So wurde in einer groß angelegten, bundesweiten Screening-Untersuchung im Rahmen der Schilddrüseninitiative „Papillon“ mittels Ultraschall an mehr als

96.000 freiwilligen berufstätigen Personen eine Knotenfrequenz von > 23% festgestellt [Reiners 2004]. Ob diese Knoten ein malignes Potenzial besitzen, ist jedoch in den seltensten Fällen mit der Sonographie allein zu lösen. Vielmehr ist es Aufgabe der Sonographie, Schilddrüsenknoten zu erkennen, volumetrisch zu bestimmen und sie bei einem Durchmesser von 10 mm und mehr einer funktionellen Charakterisierung mittels Szintigraphie zuzuführen.

Besonders sensitiv ist die Sonographie bei der Detektion nicht-palpabler knotiger Strukturen. Fujimoto et al. waren die ersten, die 1967 den Ultraschall zur Diagnostik von Schilddrüsenknoten mit einem Durchmesser > 1 cm einsetzten [Fujimoto 1967]. Mittlerweile liegt bei hochauflösenden Ultraschallgeräten die Auflösungsgrenze fokaler Läsionen bei einem Durchmesser von 2 mm [Brander 2000, Hegedüs 2003], bei großen Impedanzsprüngen sogar darunter [Iro 2000]. Sehr häufig wird die Sonographie in der Verlaufs- und Therapiekontrolle von nodösen und diffusen Strumen sowie von funktionellen Schilddrüsenerkrankungen (z.B. M. Basedow) eingesetzt [Reiners 1987a, Reiners 1987b]. Dabei werden hohe Anforderungen an Reproduzierbarkeit und Untersucherunabhängigkeit gestellt, die von der konventionellen Sonographie nur bedingt erfüllt werden.

Schon früh gab es Versuche, den Ultraschall in die Berechnung des Schilddrüsenvolumens miteinzubeziehen [Myhill 1965, Brown 1978, Tannahill 1978], da andere Verfahren wie Palpation oder Szintigraphie allein zu ungenau waren. Ein klinisches Einsatzgebiet einer möglichst exakten Schilddrüsenvolumetrie besteht darin, die Effektivität einer medikamentösen Therapie von Strumen und/oder von Schilddrüsenknoten z.B. mit Levothyroxin zu objektivieren, um so bei der Therapiekontrolle und bei Follow-up Untersuchungen Volumenänderungen nachvollziehen zu können [Cheung 1989, Reverter 1992, La Rosa 1995, Mainini 1995, Papini 1998, Zelmanovitz 1998, Wemeau 2002]. Einige neuere Metaanalysen, die die Wirkung einer Schilddrüsenhormonbehandlung auf thyreoidale Knoten untersuchten, berichten, wie schwierig es sei, frühere Studien adäquat zu bewerten und einen evidenzbasierten Therapieeffekt einer Behand-

lung mit Schilddrüsenhormonen nachzuweisen [Castro 2002, Richter 2002, Brauer 2003]. Aber schon darüber besteht keine Einigkeit, ab wann Änderungen des Knotenvolumens als relevant zu bezeichnen sind. Die willkürliche Festlegung mancher Autoren reicht von 20 - 49% [Celani 1990, Gullu 1999, Quadbeck 2002] bis hin zu 50% und mehr [Cheung 1989, Mainini 1995, Lima 1997, Wemeau 2002].

Eine weitere Domäne der Schilddrüsenvolumetrie oder der Volumenbestimmung von autonomen Adenomen ist die prätherapeutische Dosimetrie zur Ermittlung des „Zielvolumens“ vor einer geplanten Radioiodtherapie mit I-131. Hier geht die Masse des speichernden Gewebes linear in die Berechnung der benötigten I-131-Aktivitätsmenge ein [Billion 1958, Schmitz 1963]. Auch in diesem Fall ist eine hohe Messgenauigkeit erforderlich, die bei der konventionellen Sonographie allerdings nur näherungsweise gegeben ist [Wesche 1998]. Wie bei der Computertomographie gibt es zwar bestimmte Standardschnitte, aber bereits durch eine minimale Veränderung der Schallkopfposition lassen sich theoretisch unendlich viele Schnittebenen konstruieren, die hohe Ansprüche an das räumliche und anatomische Vorstellungsvermögen des Untersuchers stellen. Von mehreren Arbeitsgruppen wurde gezeigt, dass die Erfolgsquote einer Radioiodtherapie entscheidend vom prätherapeutisch definierten Zielvolumen abhängt [Peters 1995, Peters 1997, Lucas 2000, Reinhardt 2002a], das sowohl in die Berechnung der zu applizierenden Aktivitätsmenge nach der Marinelli-Formel [Marinelli 1948] als auch in die Evaluierung der erreichten Herddosis in Gray (Gy) eingeht; diese ist definiert als die Energie in Joule (J), die durch ionisierende Strahlung übertragen wird, dividiert durch das entsprechende Volumen in Milliliter (ml).

In letzter Zeit stößt die Schilddrüsenvolumetrie auf gesteigertes Interesse wegen der Einführung der minimal-invasiven Schilddrüsenchirurgie, die eine möglichst korrekte Volumenbestimmung voraussetzt. Der wesentliche limitierende Faktor bei der Auswahl von Patienten für minimal-invasive Eingriffe ist letztendlich das Volumen der Schilddrüse, das 20 ml nicht überschreiten sollte [Gagner

2001, Miccoli 2004]. Bestimmt wird das Volumen derzeit mit Hilfe einer Ultraschalluntersuchung des Halses unter Verwendung einer mathematischen Formel, in die Annahmen zur geometrischen Form der Schilddrüse bzw. von Knoten eingehen [Brunn 1981, Szebeni 1992, Hussy 2000, Shabana 2003]. Jedoch wurde die Zuverlässigkeit dieser Berechnungsmethode bisher noch nicht hinreichend genau untersucht [Hussy 2000]. Die einzige Möglichkeit, um eine korrekte Information über das Schilddrüsenvolumen zu erhalten, besteht in einem Volumenabgleich des nach totaler Thyreoidektomie gemessenen OP-Präparates mit dem präoperativ ermittelten Messwert [Schlögl 2001, Miccoli 2006, Shabana 2006].

1.3 Problem einer exakten sonographischen Schilddrüsenvolumetrie

Nach Einführung der Sonographie in die Diagnostik von Schilddrüsenerkrankungen wurden verschiedene Ultraschalltechniken sowie unterschiedliche Formeln zur Volumenberechnung vorgeschlagen und erprobt [Fujimoto 1967, Woodward 1970, Rasmussen 1974, Igl 1980, Brunn 1981, Igl 1981, Schumm 1982, Szebeni 1992]. Mittlerweile ist die Bestimmung des Schilddrüsenvolumens in vivo durch Ultraschall ein wichtiger Baustein in der klinischen Routine und stellt sowohl prä- als auch posttherapeutisch ein Standardverfahren dar [Lucas 2000, Reinartz 2002, Van Isselt 2003]. So wird die Diagnose einer Struma in erster Linie durch eine im Ultraschall nachgewiesene Volumenvergrößerung gestellt.

Wie oben bereits erwähnt, konnten mehrere Publikationen zeigen, dass der Erfolg einer Radioiodtherapie nicht nur von der Zieldosis, sondern auch vom Schilddrüsenvolumen abhängt [Peters 1995, Peters 1997, Lucas 2000, Reinhardt 2002a]. Darüber hinaus spielt die Ultraschall-Volumetrie bei großen epidemiologischen Studien eine wichtige Rolle [Brander 1989, Knudsen 2000, Reiners 2004], besonders wenn es um das Thema Iodmangelstruma geht [Vitti 1994, Knudsen 1999].

Abb.1. OP-Präparat einer irregulär konfigurierten Schilddrüse.

In der täglichen Routine wird das Volumen von Schilddrüsenlappen am häufigsten durch den zweidimensionalen (2D) Ultraschall auf der Basis des sog. konventionellen Ellipsoidmodells bestimmt [Brunn 1981, Krasznai 1985, Knudsen 1999, Van Isselt 2003]. Jedoch wird dieses mathematische Modell der irregulären Form vieler Schilddrüsen häufig nicht ganz gerecht. Andere, aufwändigere Verfahren wie das sog. „korrigierte Ellipsoidmodell“ [Szebeni 1992] oder eine Modifikation der sonographischen Volumenbestimmung unter Verwendung einer computergestützten seriellen Schnittbildtechnik [Igl 1980] erlauben zwar eine genauere Schilddrüsenvolumetrie als das konventionelle Ellipsoidmodell, konnten sich aber in der Praxis nicht durchsetzen.

Die sonographische Volumenbestimmung mit modernen Real-time Scannern basiert derzeit auf 2D-Messungen der Schilddrüsenlappen und der anschließenden Datenverarbeitung anhand der oben erwähnten geometrischen Modellvorstellung: demnach entspricht jeder Lappen einem Ellipsoid, dessen größte Durchmesser entlang seiner drei Hauptachsen angeordnet sind [Rasmussen 1974, Brunn 1981, Hegedüs 1990, Szebeni 1992]. **Richtigkeit und Präzision** dieses Modells wurden bereits früher durch Postmortem-Studien und durch Vergleichsstudien an Phantomen validiert [Szebeni 1992, Hussy 2000, Schlögl 2001, Van Isselt 2003, Schlögl 2006].

Die Volumenbestimmung mittels 2D-Ultraschall ist jedoch stark untersucherabhängig. Bei der Schilddrüsensonographie lässt sich insbesondere zwischen den Ergebnissen mehrerer Untersucher oft nur ein mäßiger Grad an Übereinstimmung feststellen [Jarløv 1993, Özgen 1999, Brauer 2005]. Glaubt man den Ausführungen von Szebeni et al. [Szebeni 1992], so ist bei dem häufig verwendeten vereinfachten Ellipsoidmodell mit einem beträchtlichen zufälligen Fehler (Random Error) zu rechnen, der bisweilen über 100% liegen kann. Daher wurde mit Nachdruck nach einer Methode gesucht, um das Volumen so exakt wie

möglich zu ermitteln [Hegedüs 1983, Szebeni 1992, Hussy 2000, Nygaard 2002, Reinartz 2002, Lyshchik 2004b]. Ziel mehrerer Untersuchungen war es dann, die Untersucherabhängigkeit als Fehlerquelle in der Schilddrüsenvolumetrie zu quantifizieren [Knudsen 1999, Özgen 1999, Peeters 2003, Wienke 2003].

Bei sonographischen Untersuchungen der Schilddrüse geht es in erster Linie darum, mittels 2D-Ultraschall das Gesamtvolumen bzw. die Volumina einzelner Schilddrüsenknoten meist im Verlauf zu bestimmen. Diese Volumenberechnungen werden in der täglichen Routine häufig von unterschiedlichen Untersuchern durchgeführt, was unterschiedliche Messergebnisse zur Folge hat. Aber auch bei einem einzelnen Untersucher ergeben sich im Laufe mehrerer Messungen diskrepante Resultate. Denn selbst in diesem Fall und bei nur einer Serie von Bildern fällt die Wahl der „korrekten“ Durchmesser zur Berechnung des Schilddrüsenvolumens oft ziemlich willkürlich aus. Da ferner jedes 2D-Bild einen Schnitt durch das Organ in einer ganz bestimmten Position und Orientierung repräsentiert, ist es nahezu unmöglich, die identische Schnittführung später zu reproduzieren, was gerade die Volumenbeurteilung im Verlauf erschwert.

Die wichtigsten Fehlerquellen dieser Methode sind somit eine erhebliche Untersucherabhängigkeit (Subjektivität) und eine eingeschränkte Reproduzierbarkeit, was sich in Messungenauigkeiten niederschlägt, die sich mit den Begriffen **Intra- und Interobserver-Variabilität** charakterisieren lassen [Jarløv 1993, Özgen 1999, Brauer 2005]. Diese stellen die wesentlichen Limitationen für die **Reliabilität (Zuverlässigkeit)** von Schilddrüsenvolumenbestimmungen in der klinischen Praxis dar.

Obgleich in vielen Studien gezeigt wurde, dass der konventionelle 2D-Ultraschall unter Verwendung der ellipsoiden Modellvorstellung methodische Schwächen aufweist [Brunn 1981, Szebeni 1992, Knudsen 1999, Schlögl 2001, Nygaard 2002, Van Isselt 2003], sind deren Ausmaß und Auswirkung auf die klinische Anwendung bisher noch nicht hinreichend genau charakteri-

siert worden. Eine wichtige Aufgabe ist es daher, **Intra- und Interobserver-Variabilität** dieser Untersuchungsmodalität und dieses Modells möglichst präzise zu bestimmen. Denn nur durch eine mathematisch exakte Quantifizierung der Einflussgrößen können die Ergebnisse von sonographischen Volumenberechnungen besser bewertet und Wege aufgezeigt werden, um **Richtigkeit und Präzision** der Schilddrüsenvolumetrie durch den 2D-Ultraschall zu verbessern. Zu diesem Zweck wurden Datensätze, die mit der 2D-Methode (konventionelles Ellipsoidmodell) an gesunden Probanden generiert wurden, mit 3D-Ultraschall-Referenzvolumina verglichen.

1.4 Definition eines validen Goldstandards

Bis heute gibt es in der bildgebenden Diagnostik keinen validierten Goldstandard zur Bestimmung des Schilddrüsenvolumens. Bisher fand man eine sehr gute Richtigkeit (accuracy) und Reproduzierbarkeit (precision) sowohl bei der Kernspintomographie (MRT) als auch bei der Computertomographie (CT) und dem 3D-Ultraschall [Schlögl 2001, Nygaard 2002, Van Isselt 2003]. Aufgrund ihres hohen, gut dokumentierten Stellenwertes in der Schilddrüsendiagnostik [Noma 1987, Huysmans 1994, Loevner 1996, Naik 1998] wird von manchen Autoren die MRT als Goldstandard für die Bestimmung des Schilddrüsenvolumens angesehen [Reinartz 2002, Van Isselt 2003].

Für die vorliegende Studie wurde als Referenzmethode die 3D-Ultraschalltechnik gewählt, da sie an Leichenschilddrüsen als dem eigentlichen **Goldstandard** evaluiert wurde [Schlögl 2001, Lyshchik 2004a] und trotz ihrer Anfälligkeit für Artefakte und qualitative Beeinträchtigungen (Minderung der Bildqualität z.B. durch willkürliche Patientenbewegungen, tiefes Ein- und Ausatmen oder Schluckakte während des Scanvorgangs [Lyshchik 2004a] keinen systematischen Fehler aufweist [Schlögl 2001]. Außerdem ist sie bzgl. Richtigkeit und Präzision bzw. Reproduzierbarkeit dem konventionellen 2D-Ultraschallverfahren klar und statistisch signifikant überlegen [Gilja 1994, Riccabona 1995, Riccabona 1996, Chang 1997, Tong 1998, Schlögl 2001, Lyshchik 2004a]. Der 3D-

Ultraschall ist ferner deutlich kostengünstiger und breiter verfügbar als die MRT oder die CT und kommt – anders als die CT – ohne Röntgenstrahlung aus.

Abb 2. Ermittlung des Schilddrüsenvolumens eines Autopsie-präparates durch die Submersionsmethode.

In der Studie von Schlögl et al. [Schlögl 2001] wurden zwei unterschiedliche 3D-Ultraschall-Methoden zur Schildddrüsenvolumenbestimmung getestet: die komplette Segmentation und die multiplanare Volumenapproximation (MVA). Um den Messfehler in der Volumenbestimmung der 3D-Methode beurteilen zu können, wurde als Referenzmethode die Volumenbestimmung von resezierten Leichenschilddrüsen durch Submersion nach dem archimedischen Prinzip der Wasserverdrängung gewählt, die den einzigen, wahren Goldstandard darstellt. Die Analyse der Daten ergab einen Mittelwert von -3,6% für die Segmentation und von -2,6% für die MVA. Die Standardabweichungen im Vergleich zur Referenzmethode betrugen 9,7% (Segmentation) und 11,5% (MVA). Diese Ergebnisse belegen, dass der systematische Fehler bei Verwendung der MVA-Methode vernachlässigbar ist.

1.5 Volumetrische Analysen mit einem neuartigen Schilddrüsenphantom

Die rasche Volumenzunahme eines hypoechogenen soliden Herdbefundes ist suspekt für ein Schilddrüsenkarzinom. Demgegenüber werden maligne Veränderungen sehr selten in echonormalen oder echodichten Läsionen gefunden [Wiedemann 1982]. Jedoch gibt es keine sicheren sonographischen Kriterien (sog. Halo-Zeichen, zystische Degeneration oder Verkalkung) für die Frühdiagnose eines Karzinoms [Solbiati 1993, Freitas 1994, Hegedüs 1998, Rago 1998], wohl aber können daraus Informationen abgeleitet werden, die eine bessere Differenzierung benigner Läsionen von malignen Herdbefunden erlauben [Lyshchik 2005]. So gehen Hegedüs und Karstrup davon aus, dass mindestens 60 – 70% der kalten solitären Knoten mittels konventioneller Sonographie und ultraschallgestützter Feinnadelbiopsie als gutartige Kolloidknoten eingestuft werden können mit einem Risiko < 1%, einen malignen Befund zu übersehen [Hegedüs 1998] .

All diese Daten belegen, wie wichtig die sonographische Erfassung und die volumetrische Bestimmung intrathyreoidaler knotiger Herdbefunde und deren Charakterisierung ist. Hilfestellung bei der Qualitätssicherung dieser Untersuchungen können – wie auch bei anderen Organen und Organsystemen – Phantom-Studien leisten. Den Angaben in der Literatur zufolge wurde zum Thema Schilddrüse erst eine einzige Phantom-Untersuchung publiziert, bei der jedoch das Schilddrüsengesamtvolumen bestimmt wurde [Szebeni 1992]. Eine aktuelle Literaturrecherche und Anfragen bei Herstellern ergaben, dass kommerziell hergestellte Schilddrüsenphantome mit „zystischen“ Läsionen von sehr geringer Echogenität erhältlich sind. Herdbefunde in der Schilddrüse sind jedoch nicht unbedingt nur zystisch. Daher sollte ein Schilddrüsenphantom neben echoarmen Läsionen auch hypoechogene Foci simulieren, die nur schwer vom umgebenden Parenchym abzugrenzen sind und somit eine diagnostische Herausforderung für den Untersucher darstellen. Anhand dieser Vorgaben wurde ein neuartiges Schilddrüsenphantom entwickelt, das hier vorgestellt wird.

Das Schilddrüsenphantom verkörpert als statisches Objekt eine untersuchungstechnische Idealsituation, kann aber die klinische Realität nicht vollständig abbilden. Anders als bei Patienten gibt es jedoch beim Phantom keine störenden Einflüsse durch Atmung und Schluckakte. Morphologisch zeigen sich keine Blutgefäße, die Schilddrüse reicht nicht nach retroklavikulär oder intrathorakal und die Läsionen bzw. Schilddrüsenlappen sind homogen und frei von Irregularitäten. Die größten Vorteile eines Phantoms liegen darin, dass es einfach zu handhaben und jederzeit verfügbar ist. Außerdem ist wegen der vorgegebenen Volumina jederzeit eine exakte Volumetrie mit genauer Reproduzierbarkeit der Volumenmessungen möglich. Die Simulation der Feinnadelbiopsie wäre zwar wünschenswert, würde aber die physische Integrität des aktuellen Phantoms irreversibel verändern.

1.6 Fragestellung und Zielsetzung

Mehrere Autoren haben schon die **Intra- und Interobserver-Variabilität** bei der Bestimmung des Schilddrüsenvolumens und knotiger Herdbefunde mit Hilfe des Ultraschalls evaluiert [Brunn 1981, Knudsen 1999, Özgen 1999, Schlögl 2001, Brauer 2005]. Darüber hinaus wurde über Interobserver-Korrelationen für Schilddrüsenvolumenmessungen berichtet [Olbricht 1983, Gutekunst 1988, Knudsen 1999]. Es gibt jedoch keine prospektive verblindete Studie, die die **Intra- bzw. Interobserver-Variabilität** bei der Volumenbestimmung der gesamten Schilddrüse an gesunden Probanden bzw. einzelner Knoten unterschiedlicher Echogenität an einem Phantom untersucht hat. Die Ergebnisse der Einzelstudien sollen hier vorgestellt und – soweit möglich – miteinander verglichen werden.

1.6.1 Phantom-Studie

Im Rahmen einer quantitativen Studie mit dem hier vorgestellten Schilddrüsenphantom soll die **Intra- und Interobserver-Variabilität** bei der 2D-Ultraschallvolumetrie einzelner Knoten unterschiedlicher Größe und Echogenität und der Schilddrüsenlappen evaluiert werden. Da Schilddrüsenknoten wegen des ge-

ringeren Volumens und ihrer oft unscharfen Randkontur schwieriger zu entdecken und auszumessen sind als die Gesamtschilddrüse, soll untersucht werden, welche Größenordnungen des Messfehlers auftreten und in welcher Relation sie zueinander stehen. Außerdem soll der methodenimmanente Fehler quantifiziert und detektierbare Volumenänderungen erfassbar gemacht werden. Bisher war in der Schilddrüsensonographie kein geeignetes Phantom verfügbar, das kommerziell erhältlich ist und mit dem qualitativ unterschiedliche intrathyreoidale Herdbefunde untersucht werden können.

1.6.2 Probanden-Studie

Die vorliegende Studie an gesunden Probanden hatte das primäre **Ziel**, die Frage nach der Quantifizierbarkeit von Unsicherheitsfaktoren in der Schilddrüsenvolumetrie durch den konventionellen 2D-Ultraschall im Vergleich zu dreidimensionalen (3D) Referenzvolumina bei gesunden Erwachsenen möglichst exakt zu beantworten und die Untersucherabhängigkeit der Methode zu demonstrieren. Damit soll die **Genauigkeit** (Richtigkeit und Präzision) der sonographischen Schilddrüsendiagnostik mathematisch erfasst und eine bessere Bewertungsgrundlage für die Frage nach der Reproduzierbarkeit von Ultraschall-Volumenbestimmungen der Schilddrüse und ihrer pathologischen Veränderungen geschaffen werden. Hierfür wurden möglichst aussagekräftige statistische Parameter wie die **Intra- und Interobserver-Variabilität**, der systematische und zufällige Fehler, der reine Fehler der Messmethode, minimale, sicher detektierbare Volumenänderungen und im Rahmen einer multivariaten Reliabilitätsanalyse die Reliabilitätskoeffizienten untersucht. Ein weiteres Ziel dieser Studie bestand darin, die **Reliabilität** der in der klinischen Routine benutzten Ellipsoidformel zur Berechnung des Schilddrüsenvolumens zu überprüfen.

2. MATERIAL UND METHODEN

2.1 Ultraschalltechnik und Volumenberechnungen

2.1.1 B-Mode-Sonographie der Schilddrüse

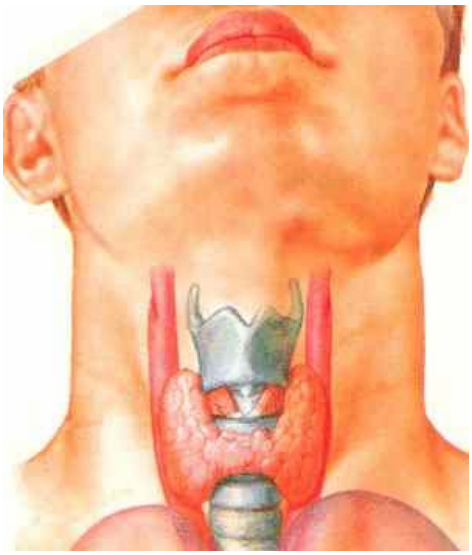


Abb.3. Topographisch-anatomische Lage der menschlichen Schilddrüse.

Zur Akquisition der planaren Ultraschallbilder wurde ein kommerzieller, hochauflösender Real-time Scanner der Firma Siemens (Elegra, Siemens AG, Medical Solutions) verwendet, der mit einem 5,0 MHz- (gekrümmt) bzw. einem 7,5 MHz- (linear) Schallkopf und einem direkten Bildspeicher (Cineloop), aus dem die Bilder zurückverfolgt werden konnten, ausgestattet ist. Dies entspricht den Empfehlungen in der Leitlinie zur Schilddrüsendiagnostik der Deutschen Gesellschaft für Nuklearmedizin [Dietlein 2003].

Ultraschall unterliegt, wie andere Schallfrequenzbereiche, in der Diagnostik denselben physikalischen Zusammenhängen. Seine Anwendung ist dadurch erweitert, dass diese Schwingungen optisch sichtbar gemacht werden können und aufgrund ihrer kleinen Wellenlänge gut zu bündeln und damit gezielt auszurichten sind.

Die Grundlagen des hier verwendeten **B-Modes** (Brightness) stellen die verschiedenen Grauwerte entsprechend der Höhe der Energie der Echos als eigentliche Information dar. Die Zuordnung der Objekttiefe erfolgt dabei über die Messung der Laufzeit der Ultraschallwellen im Gewebe [Iro 2000]. Die piezoelektrischen Signale werden im B-Mode-Verfahren durch Helligkeitsmodulation dargestellt. Die abgebildeten Echos (Bildechos) sind nicht mehr identisch mit den ursprünglichen Echosignalen (Ultraschall). Sie stellen aber eine repräsentative Projektion der reflektierten akustischen Echosignale dar. Von allen sonographischen Schnittbildverfahren hat die Sonographie im B-Bild das höchste Auflösungsvermögen, was sie für die Untersuchung der Halsweichteile besonders qualifiziert. Dem erfahrenen Untersucher liefert sie ein maßstabgetreues, zweidimensionales Schnittbild mit hohem Informationsgehalt, wobei durch „Abscannen“ einer Region sich im Verlauf der dynamischen Untersuchung ein räumlicher Eindruck ergibt. Für die Untersuchung ist, neben profunden anatomischen Kenntnissen, auch ein gutes räumliches Vorstellungsvermögen wichtig. Neben den technischen und praktischen Voraussetzungen bestimmen sie die Qualität und Treffsicherheit der Untersuchungsmethode.

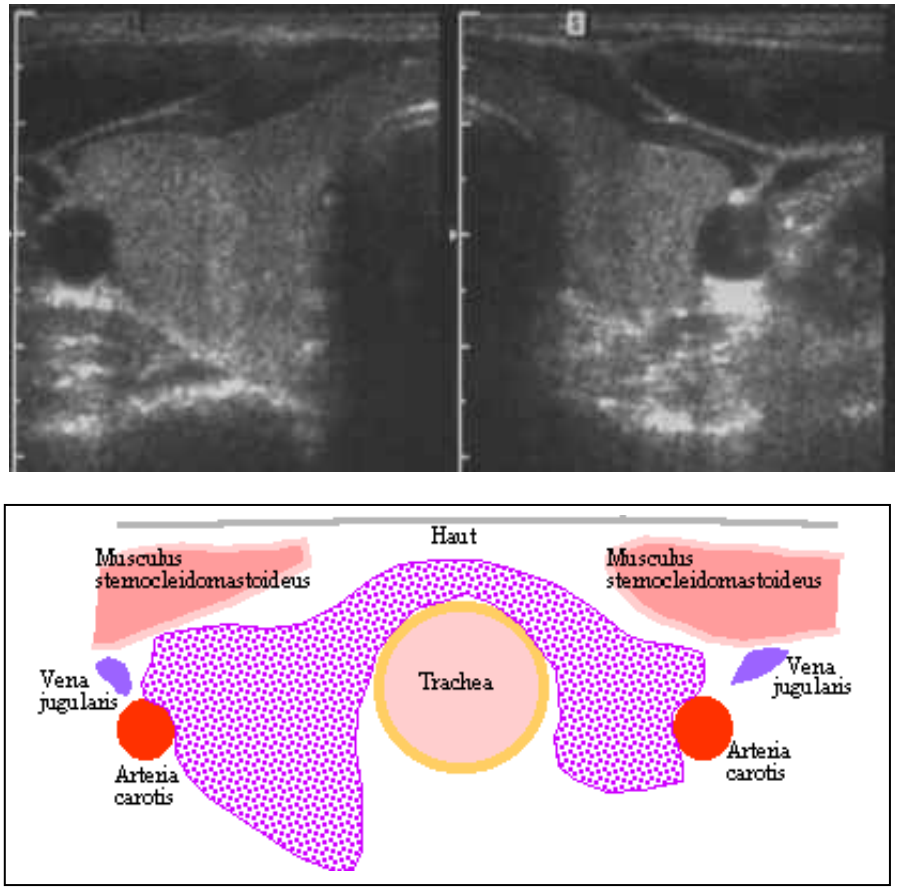


Abb. 4. Sonographischer Querschnitt durch eine normal konfigurierte Schilddrüse. Typischerweise stellt sich das Parenchym mit einer homogenen, fein granulierten Echostruktur dar. Im Vergleich zur umgebenden Halsmuskulatur ist das Binnenreflexmuster der Schilddrüse echoreicher.

Um mit der Simulation der alltäglichen klinischen Untersuchungsroutine eine möglichst weitreichende Relevanz der Ergebnisse zu erzielen, sollten bestimmte Untersuchungsmodalitäten frei wählbar sein: bei jeder Messung konnten die Untersucher zwischen dem linearen und dem gekrümmten Schallkopf wählen und waren nicht an eine bestimmte Schallkopfposition, Angulierung oder Druckausübung gebunden. Außerdem sollten keine Messungen wegen schlechter Bildqualität verworfen werden, wodurch ein statistischer bzw. Selektionsfehler vermieden wurde.

2.1.2 2D-Ultraschallvolumetrie nach der konventionellen Ellipsoidformel

Zur Bestimmung der Intra- und Interobserver-Variabilität vermaßen neun Ärzte mit unterschiedlich langer Erfahrung in der Ultraschalltechnik (3 bis 10 Jahre) die Durchmesser der Schilddrüsenlappen und -läsionen. Die Untersucher bestimmten die Lappen- und Herdvolumina je dreimal hintereinander mit einem zeitlichen Abstand von mehr als einer Woche zwischen zwei Sitzungen, um untersucherspezifische Einflüsse so gering wie möglich zu halten. Das 2D-Volumen der Schilddrüsenlappen in vivo bzw. der Knoten- und Lappenvolumina des Phantoms wurde durch Messung der durchschnittlichen Länge (A), Breite (B) und Tiefe (C) eines Objekts auf zwei orthogonalen Ultraschallbildern (longitudinale und transversale Querschnitte) und Volumenberechnung (V) gemäß der modifizierten Formel für ein Ellipsoid ermittelt:

$$V_{ellipsoid} = A\ [cm] \cdot B\ [cm] \cdot C\ [cm] \cdot f \qquad (f\text{: Koeffizient})$$

Um die Rechenoperationen zu vereinfachen und die klinische Routine zu simulieren, wurde f = 0,5 anstelle des Standardkoeffizienten $\frac{\pi}{6}$ (~0,524) [Igl 1981] für die Volumenbestimmung eines Ellipsoids bzw. dem von Brunn et al. [Brunn 1981] vorgeschlagenen, empirisch ermittelten Korrekturfaktor f = 0,479 verwendet. Das Schilddrüsengesamtvolumen wurde definiert als die Summe der Lappenvolumina. Da das Volumen des Schilddrüsenisthmus normalerweise vernachlässigbar klein ist (ca. 5%) [Szebeni 1992], wurde er nicht in die Volumenberechnung mit einbezogen.

2.1.3 Gewählter Goldstandard: 3D multiplanare Volumenapproximation

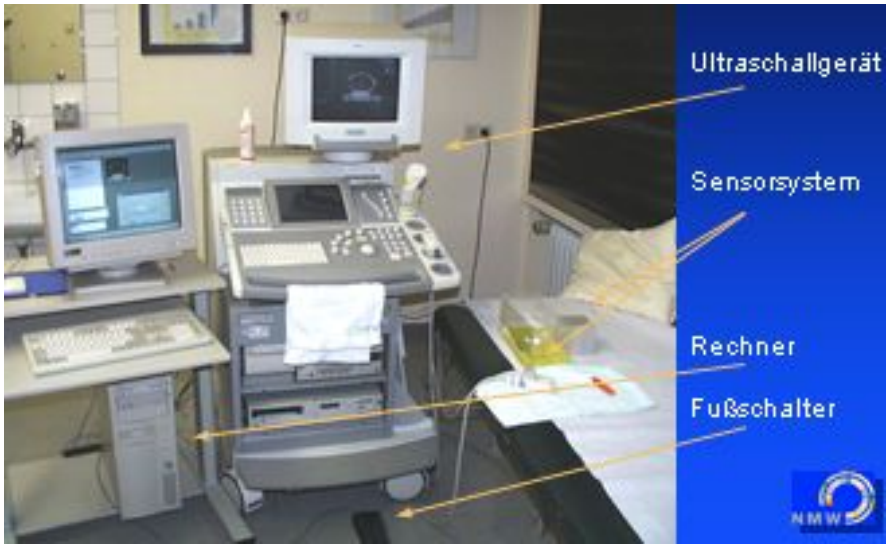


Abb. 5. Ultraschall-Arbeitsplatz in der Klinik für Nuklearmedizin der Universität Würzburg mit angeschlossenem Akquisitionsrechner für die 3D-Datensätze.

Die Schilddrüsenreferenzvolumina wurden mit Hilfe der multiplanaren Volumenapproximationsmethode (MVA) [Schlögl 2001] als dem hier gewählten Goldstandard erzeugt. Diesem Verfahren liegt eine in der Echokardiographie angewandte Methode zur Bestimmung des linksventrikulären Volumens aus zwei aufeinander senkrecht stehenden Längsachsenschnitten zugrunde. Unmittelbar nach Akquisition der 2D-Daten wird hierbei das 3D-Bild rekonstruiert. Die 3D-Ultraschallbilder werden mit dem kommerziell erhältlichen FreeScan-System (EchoTech 3D Imaging Systems, Hallbergmoss, Deutschland) generiert. Dieses besteht aus einem elektromagnetischen Sensorsystem, das an den Ultraschallkopf gekoppelt ist, einem handelsüblichen Computer (Pentium 4; 2,60 GHz) zur Bildspeicherung, -nachverarbeitung und digitalen Bildarchivierung kombiniert mit einem Framegrabber, der die analogen Videosignale digitalisiert, und der Software „FreeScan“.

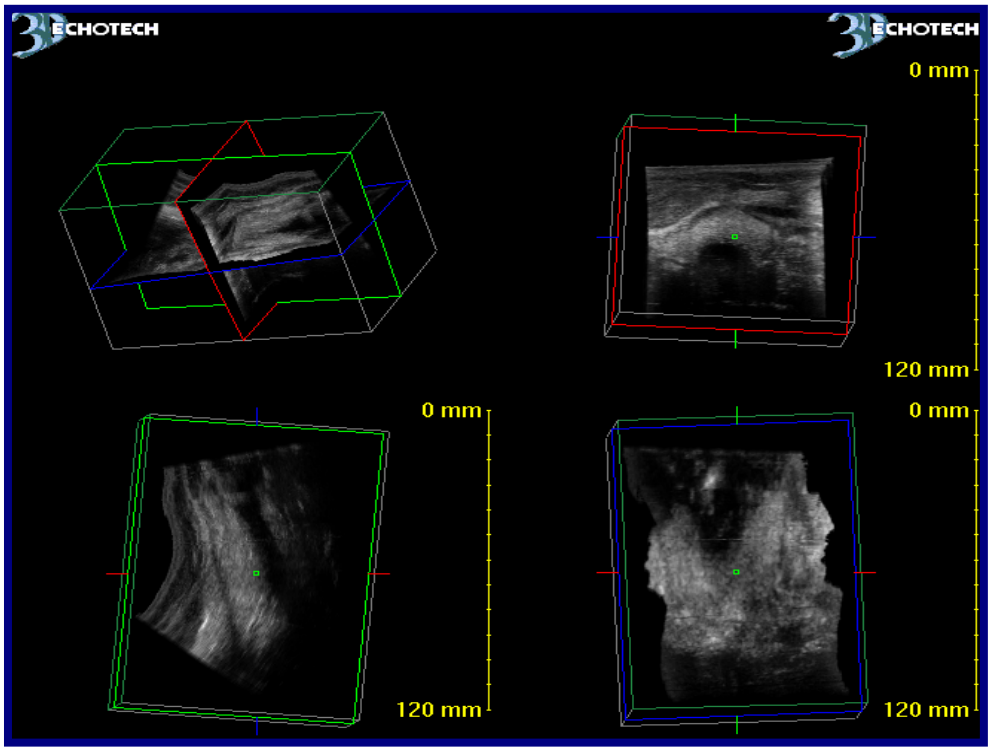


Abb. 6. 3D-Rekonstruktion von 2D-Ultraschall-Daten mittels EchoTech-Software.

Zunächst werden einmalig für jeden Ultraschallkopf, der für 3D-Akquisitionen verwendet wird, Kalibrierungsmessungen durchgeführt. Damit wird die Lage des Sensors auf dem Schallkopf relativ zum Entstehungsort der Bilder definiert, so dass für jedes Bild, das auf den Rechner übertragen wird, Ort und Lage im Feld des Transmitters festliegen. Für die 3D-Akquisition wird der Transmitter neben dem Hals der Versuchsperson positioniert und erzeugt ein elektromagnetisches Feld. Der an den Signalumwandler (Ultraschallkopf) gekoppelte Sensor übermittelt seine Position innerhalb des elektromagnetischen Feldes an einen Prozessor. Danach kann der Schallkopf wie gewohnt geführt werden. Die Bilder des Ultraschallgerätes werden als S-VHS-Videosignale übertragen und vom Framegrabber digitalisiert. Im nächsten Schritt werden die Daten vom Auswerteprogramm nachverarbeitet. Erzeugt wird ein „geordneter“ 3D-Datensatz, indem die Raumkoordinaten des Schallkopfs zusammen mit den zweidimensionalen Informationen der konventionellen Ultraschallbilder gespeichert und in eine reguläre 3D-Voxelmatrix konvertiert werden. Mehrfach erfasste Punkte werden gemittelt, nicht erfasste interpoliert. Die Abmessungen des entstehenden Quaders resultieren aus der gescannten Fläche und der Eindringtie-

fe. Dieser nachverarbeitete Datensatz stellt die Basis für die weitere Bildverarbeitung dar. Zur Verfügung stehen mehrere Messfunktionen (Entfernung, Winkel, Volumina, etc.) und Darstellungsarten (Oberflächen, Transparentdarstellungen, Animation, etc.). Es können auch nur Teile des Datensatzes (Region of Interest) nachverarbeitet werden, z.B. bei der Volumenberechnung eines Knotens.

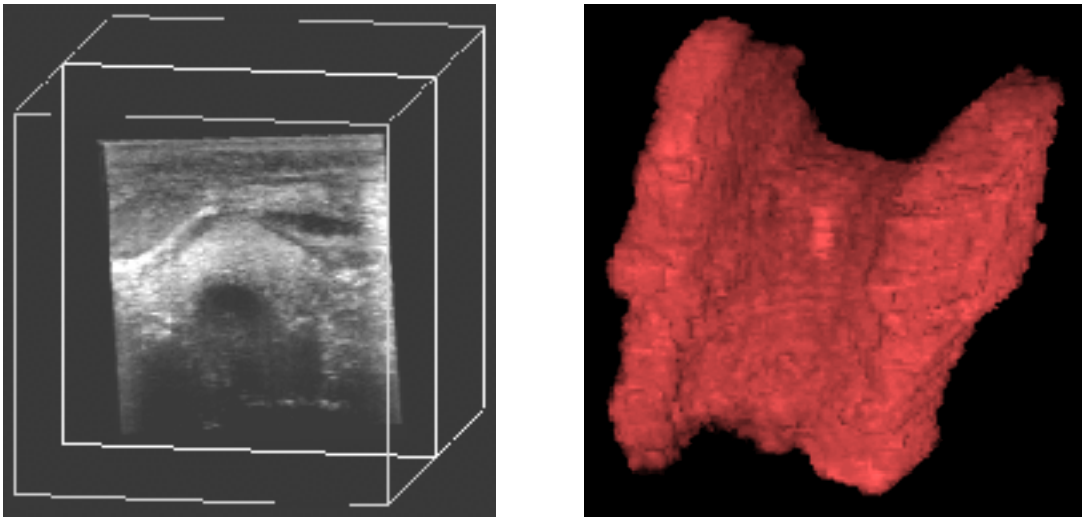


Abb.7. Nachbearbeiteter Volumenwürfel mit 2D-Schnittbild eines Schilddrüsenlappens (links). 3D-Rekonstruktion einer Schilddrüse (rechts).

Nach Auswahl der optimalen Schnittbilder wird eine Längsachse der gespeicherten Schilddrüsenlappen als Rotationsachse innerhalb des rekonstruierten 3D-Datenwürfels definiert. Die Software generiert bis zu neun planare Schnitte, die kreisförmig in festem Winkelabstand um diese Rotationsachse herum angeordnet sind. Die Organgrenzen müssen auf diesen longitudinalen Flächen vom Untersucher manuell eingezeichnet werden. Das Programm erstellt dann äquidistante Flächen, die innerhalb der vorgegebenen Konturen orthogonal zu den Segmenten angeordnet sind. Dadurch erhält man Konturpunkte, die durch eine Spline-Interpolation dritten Grades zu einer Konturlinie verbunden werden. Der Abstand zwischen den Flächen wird vom Programm möglichst klein gewählt, um eine optimale Volumenapproximation zu erreichen. Das Volumen errechnet

sich aus der Summe der eingeschlossenen Flächen multipliziert mit den Abständen zwischen den Flächen der Längsachse (Abb. 8). Diese Berechnung ist automatisiert und reduziert so den Zeitaufwand für den Untersucher. Die 3D-Volumenbilder können dann interaktiv auf einem Computermonitor betrachtet werden. Im Vergleich zur konventionellen 2D-Technik ist die 3D-Methode nicht auf Modellvorstellungen angewiesen.

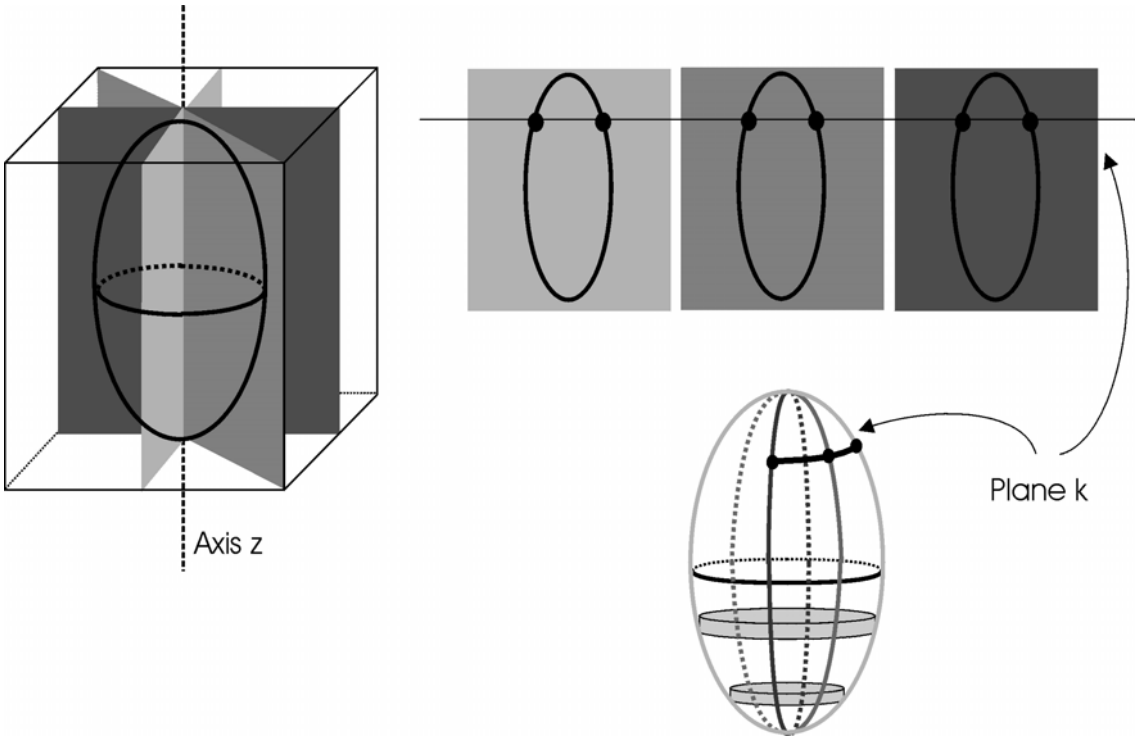


Abb. 8. Prinzip der Volumenberechnung mit der multiplanaren Volumenapproximation.

2.2 Phantom-Studie

2.2.1 Aufbau des Phantoms

Für das Phantom wurde muskeläquivalentes Gewebe verwendet, das in eine 16 x 16 x 10 cm große Box eingebracht wurde. In das Material eingebettet sind zwei Ellipsoide, die die Schilddrüsenlappen darstellen, mit Längsachsen von jeweils 45 mm und je zwei Kurzachsen von 25 mm. Daraus ergibt sich für beide Ellipsoide ein Gesamtvolumen von jeweils 14,7 ml. Im Ultraschall verhalten sich die Ausbreitungsgeschwindigkeit der Schallwellen sowie Echogenität, Schwächung und Rückstreuung beider Lappen äquivalent zu normalem Schilddrüsengewebe. Die Längsachsen beider Lappen sind in einem Winkel von 30° zueinander angeordnet und simulieren so die anatomische Lagebeziehung von Schilddrüsenlappen (Abb. 9). Auf eine Darstellung des Schilddrüsenisthmus wurde verzichtet.

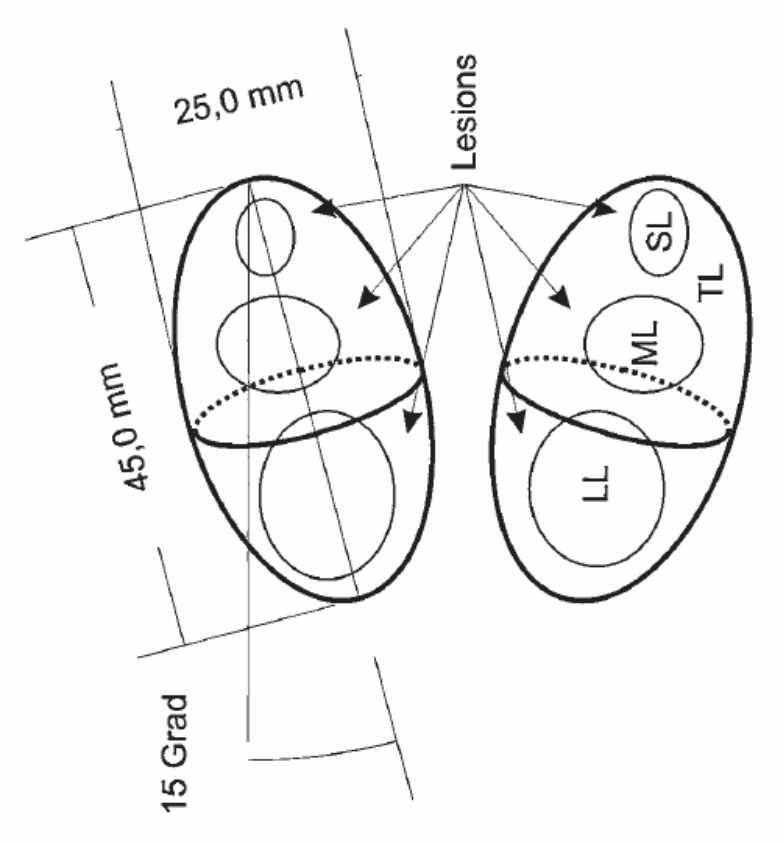


Abb. 9. Schematische Darstellung des Schilddrüsenphantoms.

Entlang der Längsachse jedes Lappens wurden drei hypoechogene ellipsoide knotige Läsionen mit unterschiedlichen Volumina angeordnet: großer Knoten = 1,4 ml, mittlerer Knoten = 0,5 ml, kleiner Knoten = 0,1 ml. Knotige Veränderungen von 0,1 ml bis 1,4 ml sind häufige Befunde bei Patienten. Da das Phantom ein statisches Objekt ist, sind die Untersuchungsbedingungen für Herdbefunde deutlich besser als bei In-vivo-Untersuchungen. Daher wurde die Größe der Läsionen so gewählt, dass die 2D-Messungen mögliche Fehler bei der Volumenbestimmung aufdecken; demzufolge sind sie auch kleiner als die typischen Befunde, die man bei Patienten findet. Zur Simulation größerer Volumina lassen sich die ganzen Lappen heranziehen. Die Längsachsen der Läsionen divergieren von der Längsachse der Lappen, um die Morphologie humaner Schilddrüsen möglichst realitätsnah abzubilden. Die Rückstreuung der Schilddrüsenobjekte liegt bei 8,5 dB. In einem Lappen beträgt die Echodifferenz der Läsionen -5 dB in Relation zum umgebenden Material, was im Ultraschall ziemlich schwierig zu differenzieren ist. Im anderen Lappen haben die Herdbefunde eine Echodifferenz von -10 dB, was typisch ist für Zysten oder andere Läsionen mit sehr geringer Echogenität.

In Abb. 10 sind zweidimensionale Ultraschall-Schnittbilder des Phantoms in Längs- und Querorientierung dargestellt. Abb. 11 zeigt ein dreidimensionales Ultraschallbild des Phantoms, wobei die -5 dB-Läsionen auf der rechten Seite und die -10 dB-Läsionen auf der linken Seite im oberen rechten Quadranten der Figur dargestellt sind. Hergestellt wurde das Phantom von der Firma Dansk Fantom Service (Jyllinge, Dänemark).

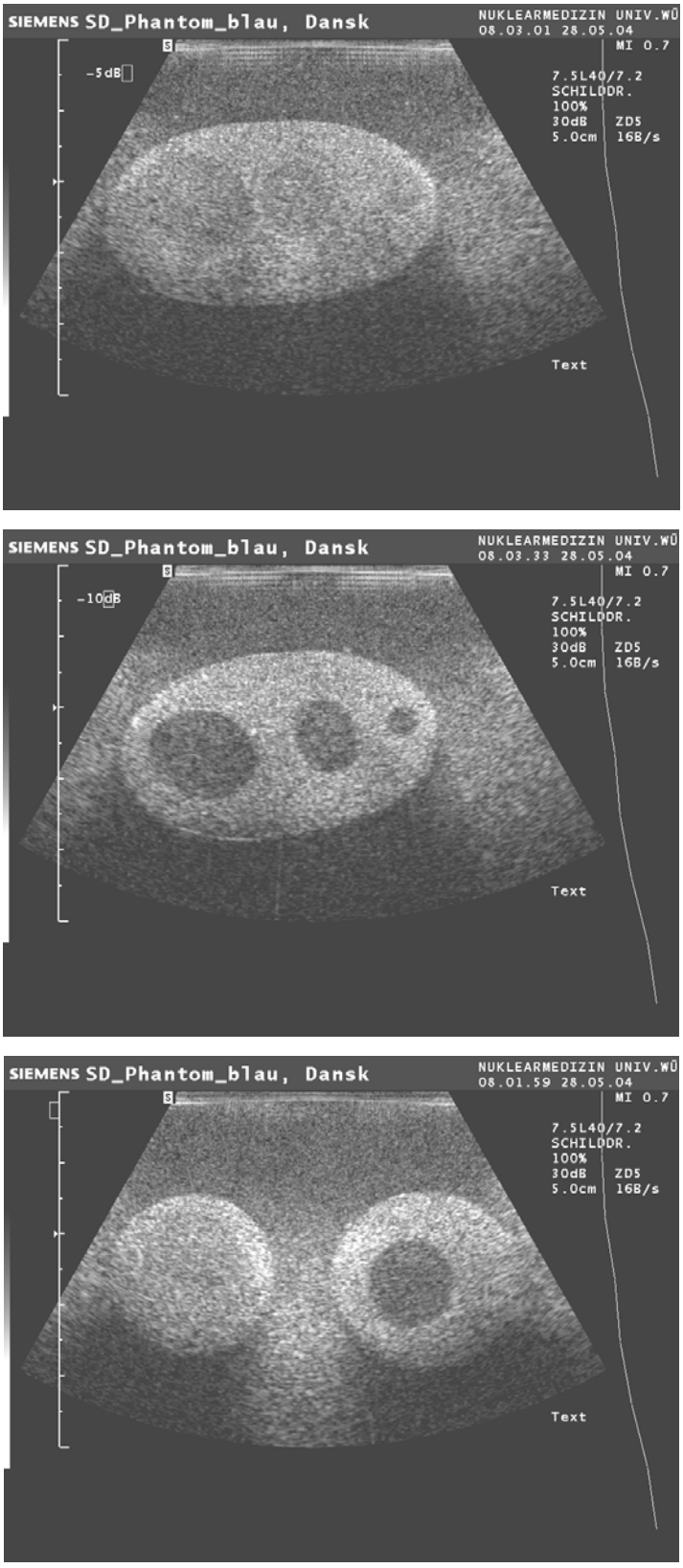


Abb. 10. Zweidimensionale Längs- und Querschnitte durch die Lappen des Schilddrüsenphantoms, in die Herdbefunde unterschiedlicher Größe und Echogenität eingearbeitet sind.

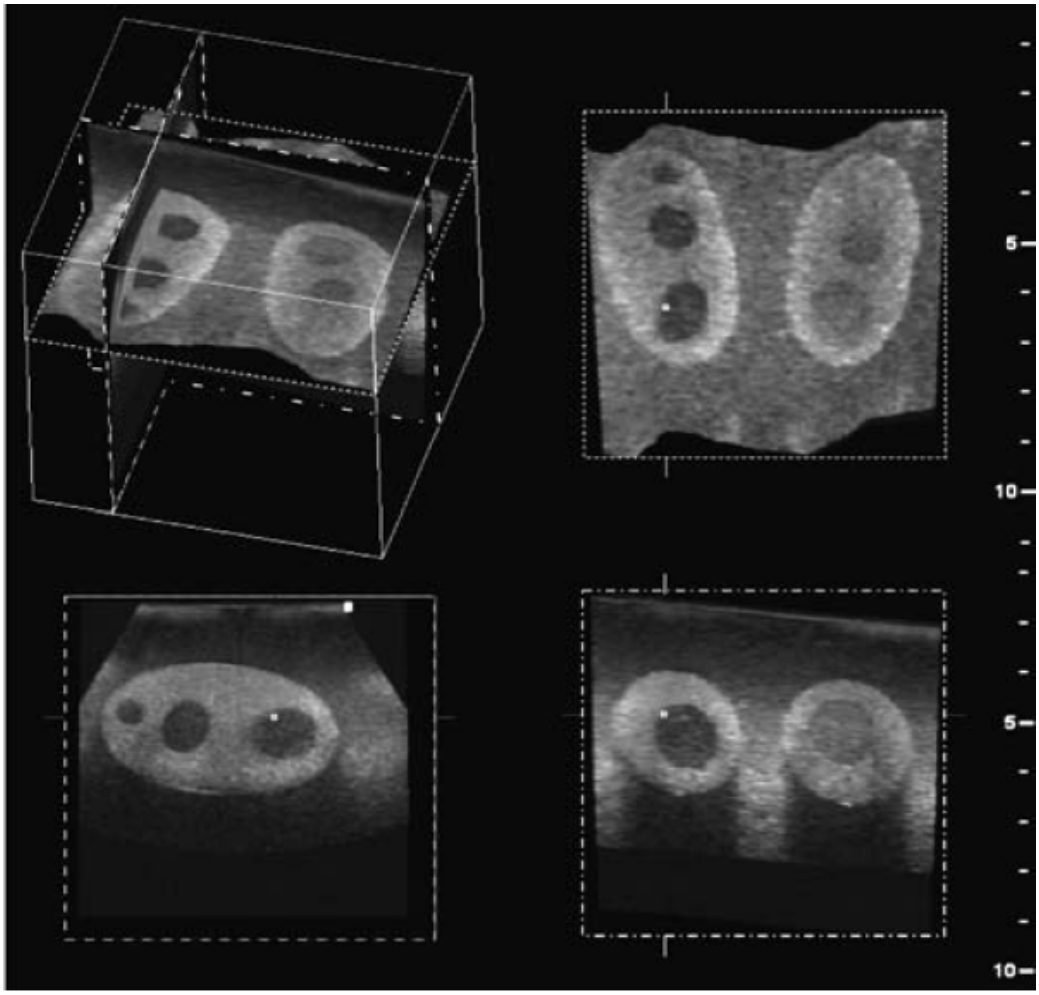


Abb. 11. Darstellung der Lappen des Schilddrüsenphantoms im 3D-Ultraschall.

2.2.2 Datenerhebung und -analyse

Die Volumina der im Phantom modellierten Lappen und Läsionen wurden dreimal pro Untersucher mittels konventionellem planaren 2D-Ultraschall nach der Modellvorstellung eines Ellipsoids (Ellipsoidmethode) bestimmt (siehe Punkt 2.1.2). Hierfür wurde wieder die Länge der drei Hauptachsen ermittelt und das Volumen nach der o.g. Ellipsoidformel berechnet. Auch bei den Volumenberechnungen am Phantom wurde der Koeffizient f = 0,5 favorisiert und bei der nachfolgenden Analyse verwendet. Pro Sitzung wurden demnach 8 Volumenberechnungen pro Untersucher durchgeführt (je 3 Läsionen mit -5 dB, 3 Läsionen mit -10 dB und 2 gesamte Lappen), was eine Gesamtzahl von 216 Scans ergab. Die Volumina aller Läsionen und Lappen wurden durch einen erfahrenen

Untersucher nach der Methode der multiplanaren Volumenapproximation bestimmt [Schlögl 2001]. Die gemessenen Volumina wurden dann mit den wahren Volumina verglichen, die vom Hersteller vorgegeben waren. Die Datensätze wurden entsprechend der Echogenität der Läsionen in zwei Subgruppen unterteilt, die zur Bestimmung der Intraobserver-Variabilität herangezogen wurden. Damit sollte die Hypothese überprüft werden, dass die Volumenbestimmung durch den Kontrast beeinflusst wird, der durch die unterschiedliche Echogenität zustande kommt.

2.3. Probanden-Studie

2.3.1 Zusammensetzung des Probandenkollektivs

Zehn gesunde erwachsene Probanden (4 Frauen, 6 Männer) im Alter zwischen 19 und 49 Jahren nahmen an der Studie teil. Alle Probanden sollten keine schilddrüsenassoziierten Symptome haben und keine oder nur sehr kleine Schilddrüsenknoten (Volumen < 1 ml) aufweisen. Untersucht wurden gesunde Erwachsene, um den möglicherweise durch Schilddrüsenpathologien induzierten Bias auf die Schilddrüsenmorphologie zu minimieren.

2.3.2 Datenerhebung und -analyse

Die Schilddrüsenvolumina aller zehn Probanden wurden je dreimal von neun mehr oder weniger erfahrenen Untersuchern mit einem jeweiligen Abstand von mindestens zwei Monaten mittels konventioneller 2D-Ultraschallvolumetrie nach der modifizierten Ellipsoidformel (vgl. Punkt 2.1.2) bestimmt, ohne vorher die wahren Volumina zu kennen. Zusätzlich wurde bei der ersten Messung ein 3D-Datensatz pro Schilddrüsenlappen generiert, um für jeden Probanden ein Referenzvolumen zu definieren. Für die 2D- und 3D-Messung gab es kein vorgegebenes Ablaufprotokoll. Alle Probanden wurden in Rückenlage mit leicht reklinierter Kopfhaltung untersucht. Das Schilddrüsenvolumen wurde für jeden Lappen getrennt berechnet. Während der Datenerhebung hatten die Untersucher keine Kenntnis von den Resultaten der anderen Untersucher.

2.4 Statistische Methoden

2.4.1 Richtigkeit und Präzision

Die **Richtigkeit (accuracy)** gibt Auskunft darüber, wie weit ein ermittelter Wert vom anerkannten Referenzwert (manchmal als wahrer Wert bezeichnet) abweicht bzw. abweichen kann. Ein Experiment kann präzise, also reproduzierbar, aber ungenau sein. Die Richtigkeit ist also ein Maß für die Abweichung eines Messwertes aufgrund eines systematischen Fehlers. Das Fehlen von systematischen Fehlern ist somit eine Grundvoraussetzung für die Richtigkeit. Dies sagt jedoch nichts darüber aus, wie stark einzelne Messwerte zufällig streuen können.

Die **Präzision (precision)** ist ein Maß für die Übereinstimmung zwischen unabhängigen Messergebnissen unter festen Bedingungen und damit für den Grad der Reproduzierbarkeit eines Experiments. Liegen mehrere Messwerte dicht beieinander, so hat die Messmethode eine hohe Präzision. Die Präzision beschreibt also die zufällige Streuung von Messwerten (zufälliger Fehler). Das bedeutet aber noch nicht, dass die gemessenen Werte auch richtig sind.

Überträgt man diese Unterscheidung zwischen Richtigkeit und Präzision auf die Statistik, so liegt es auf der Hand, dass die Richtigkeit einer Messung durch den Mittelwert (oder allgemein irgendein Lagemaß der Messwerteverteilung) beschrieben werden kann, die Präzision durch die Standardabweichung (oder durch ein anderes Streumaß).

Der Begriff Genauigkeit wird häufig mit Präzision gleichgesetzt. Die Genauigkeit ist jedoch kein Validierungselement, sondern der Oberbegriff für Richtigkeit und Präzision und ein Maß für die Übereinstimmung zwischen dem (einzelnen) Messergebnis und dem wahren Wert der Messgröße. Sie wird auch durch den Begriff Variabilität ausgedrückt. Eine hohe Genauigkeit ist nur zu erreichen, wenn sowohl die Richtigkeit als auch die Präzision gut sind. Ein Ergebnis ist somit genau, wenn es frei von zufälligen und systematischen Fehlern ist.

Die wahre Genauigkeit eines Verfahrens lässt sich nur bestimmen, wenn die exakten zu messenden Größen bekannt sind. Weil diese auch mit irgendeiner Methode gemessen werden müssen, kann genau genommen immer nur von einem Vergleich von Messverfahren gesprochen werden. Im Idealfall gibt es eine Referenzmethode (Goldstandard), deren Fehler gegenüber dem zu untersuchenden Verfahren vernachlässigbar ist. Gibt es solch ein Verfahren nicht, müssen die Fehlerquellen beider Methoden in die Genauigkeitsanalyse einbezogen werden.
Bei dem hier untersuchten 2D-Ultraschallverfahren liegen als Referenzwerte bei den Phantomobjekten die aus den geometrischen Eigenschaften sehr genau bestimmten Volumina vor. Bei den Schilddrüsenmessungen an Probanden dient das aufwändige 3D-Ultraschallverfahren als Referenzmethode; deren Fehler ist im Vergleich zu der 2D-Ultraschallmethode sehr gering [Gilja 1994, Riccabona 1995, Riccabona 1996, Chang 1997, Schlögl 2001, Lyshchik 2004a].

Die Fehler eines Messverfahrens hängen im Allgemeinen von verschiedenen Faktoren ab. Bei den hier untersuchten Verfahren werden gewisse Objekte von Untersuchern mittels einer Messapparatur gemessen. Die Genauigkeit des Messresultats hängt also (1) vom Messgerät, (2) vom Untersucher und (3) vom Objekt ab. Während der zufällige Messfehler des Gerätes konstruktionsbedingt festliegt, lässt sich u. U. der Untersucherfehler etwa durch Schulung der Untersucher reduzieren. Bei dem hier untersuchten Verfahren entstehen die Messwerte durch Interaktion des Untersuchers mit dem (bildgebenden) Gerät, d.h. der apparate- und der untersucherbedingte Messfehler lassen sich nicht separat bestimmen. Wenn im Folgenden vom methodenimmanenten oder reinen Messfehler die Rede ist, ist genau genommen immer der Gerätefehler (Bildauflösung usw.) zusammen mit einem mittleren Untersucherfehler pro Messobjekt gemeint. In der Literatur ist dafür auch der Begriff "Inner-Subjekt-Fehler" anzutreffen.

Weiterhin muss unterschieden werden, ob ein oder mehrere Untersucher Messungen vornehmen. Im ersten Fall, spricht man von **Intraobserver-Variabilität**, andernfalls von **Interobserver-Variabilität**.

Ein weiterer Aspekt bei der Beurteilung von Messverfahren ist das der **Zuverlässigkeit (Reliabilität)**, Konsistenz oder Reproduzierbarkeit. Der Reliabilitätskoeffizient ist als Anteil der Untersuchervarianz an der Gesamtvarianz definiert. Für die Berechnung ist somit kein Referenzverfahren nötig.

2.4.2 Intra- und Interobserver-Variabilität

In der Literatur sind verschiedene Maße zur Kennzeichnung von Inter- und Intraobserver-Variabilität anzutreffen. Die Definition hier orientiert sich an der klinischen Praxis, bei der es oft darauf ankommt, Größen*veränderungen* am gleichen Subjekt zu verschiedenen Zeitpunkten zu entdecken. Die Intraobserver-Variabilität beschreibt dann den Fall, in dem der gleiche Untersucher, die Interobserver-Variabilität, in dem verschiedene Untersucher die Messungen vornehmen. In beiden Fällen werden die Messungen zwar am gleichen Subjekt (Patient oder Proband), aber genau genommen an verschiedenen Objekten (Schilddrüse zu verschiedenen Zeitpunkten) vorgenommen, d.h. es muss der Messfehler zwischen den Objekten berücksichtigt werden.

Im vorliegenden Design messen r Untersucher n Objekte je m-mal. Von den $r \cdot n \cdot m$ Messwerten werden die jeweiligen Referenzwerte abgezogen. Bezieht man diese Differenzen auf den jeweiligen Referenzwert (relative Differenzen), werden die Messfehler weitgehend unabhängig von der Objektgröße. Mittelwert und Varianz aller $r \cdot n \cdot m$ Differenzen beschreiben dann die Richtigkeit und Präzision der Methode, also die Interobserver-Variabilität.

Der Intraobserver-Fehler pro Untersucher wird durch Mittelwert und Varianz aller $n \cdot m$ Werte dieses Untersuchers geschätzt. Können die Untersucher als zufällig ausgewählt betrachtet werden, repräsentiert der Mittelwert der r Varian-

zen die Intraobserver-Varianz der Methode. Schließlich ist der Mittelwert über alle $r \cdot n$ Objekt-Varianzen ein Maß für die reine Messvarianz oder Inner-Subjekt-Varianz σ_{err}^2.
Können diese Werte nicht über alle Objekte geschätzt werden, wie etwa bei der Phantom-Studie, bei denen ausgesuchte Objekte gewisse Größenklassen repräsentieren, sollten die Fehler objektbezogen oder gepoolt angegeben werden.

In Formeln:
Es bezeichne x_{ijk} den k-ten Wert ($k = 1…m$) des j-ten Untersuchers ($j = 1…r$) am i-ten Objekt ($i = 1…n$). Definiert werden

$$\bar{x} = \frac{1}{n \cdot r \cdot m} \sum_{i,j,k} x_{ijk} \;, \quad s^2 = \frac{1}{n \cdot r \cdot m - 1} \sum_{i,jk} (x_{ijk} - \bar{x})^2$$

als Gesamt-Mittelwert und –Varianz,

$$\bar{x}_{ij} = \frac{1}{m} \sum_{k} x_{ijk} \;, \quad s_{ij}^2 = \frac{1}{m - 1} \sum_{k} (x_{ijk} - \bar{x}_{ij})^2$$

als Mittelwert und Varianz pro Objekt i und Untersucher j (Inner-Subjekt) und

$$\bar{x}_j = \frac{1}{n \cdot m} \sum_{i,k} x_{ijk} \;, \quad s_j^2 = \frac{1}{n \cdot m - 1} \sum_{i,k} (x_{ijk} - \bar{x}_j)^2$$

als Mittelwert und Varianz für Untersucher j.

Dann werden Bias und Varianz jeweils geschätzt durch

$$\hat{\mu}_{inter} := \bar{x} \quad \text{und} \quad \hat{\sigma}_{inter}^2 := s^2 \quad \text{(Interobserver)}$$

und

$$\hat{\mu}_{intra,j} := \bar{x}_j \quad \text{und} \quad \hat{\sigma}_{intra,j}^2 := s_j^2 \quad \text{(Intraobserver für Untersucher } j\text{)}.$$

Gemittelt über alle Untersucher schätzen

$$\hat{\mu}_{intra} := \frac{1}{r}\sum_j \bar{x}_j \quad (= \bar{\bar{x}} = \hat{\mu}_{inter}) \quad \text{und} \quad \hat{\sigma}^2_{intra} := \frac{1}{r}\sum_j s_j^2$$

Intraobserver-Bias und -Varianz der Methode.

Die Messvarianz oder Inner-Subjekt-Varianz ist schließlich $\hat{\sigma}^2_{err} := \frac{1}{n \cdot r}\sum_{i,j} s_{ij}^2$.

2.4.2.1 „Systematic observer error“ und „random observer error“

Aus Gründen der Vergleichbarkeit mit der Literatur sind im Ergebnisteil noch die von Tong et al. [Tong 1998] eingeführten Begriffe des systematischen und zufälligen Untersucherfehlers („systematic observer error“ und „random observer error“) aufgeführt. „Systematic observer error“ ist ein Maß, mit dem die Richtigkeit der Messungen *eines* Untersuchers beschrieben wird und ein Synonym für den Intraobserver-Bias, d.h. die mittlere Abweichung der Messungen pro Untersucher über alle Objekte. Unter dem „random observer error“ verstehen Tong et al. die mittlere untersucherbezogene Inner-Subjekt-Streuung.

2.4.2.2 Inner-Subjekt-Variationskoeffizient CV

Gelegentlich wird in der Literatur auch der sogenannte Inner-Subjekt-Variationskoeffizient *CV* (Inner Subject Coefficient of Variation) als Maß für die Variabilität einer Messmethode angegeben. Es handelt sich dabei um einen mittleren Variationskoeffizienten ($\frac{Streuung}{Mittelwert}$). Diese Größe kann auf verschiedene Weise definiert sein – in jedem Fall sollte aber über die Quadrate der objektbezogenen (Inner-Subjekt) Variationskoeffizienten gemittelt und dann die Wurzel gebildet werden, andernfalls resultieren deutlich zu kleine Werte [Bland 2006].

Wird jedes Objekt *i* von Untersucher *j* mit jeder Methode genau einmal gemessen (x_{ij1}, x_{ij2}), so ist der Inner-Subjekt-Variationskoeffizient CV_{ij} definiert durch

$$CV_{ij} = \sqrt{\frac{VAR(x_{ij})}{\overline{x}_{ij}^2}} = \sqrt{2 \cdot \frac{(x_{ij1} - x_{ij2})^2}{(x_{ij1} + x_{ij2})^2}} = \sqrt{2} \cdot \frac{\left|x_{ij1} - x_{ij2}\right|}{x_{ij1} + x_{ij2}}$$

und der Inner-Subjekt-Variationskoeffizient der Methode

$$CV = \sqrt{\frac{1}{n \cdot r} \sum_{ij} CV_{ij}^2} \; .$$

Sind die Messungen auf Referenzwerte x_{ref} normiert, stimmt der objektbezogene Variationskoeffizient ungefähr mit der oben definierten Inner-Subjekt-Streuung s_{ij} überein. Dann ist nämlich

$$s_{ij} = \sqrt{\frac{VAR(x_{ij})}{x_{ref,i}^2}} \; ,$$

wenn x_{ijk} hier für die absoluten Messwerte stehen.

Der Inner-Subjekt-Variationskoeffizient wird häufig als Maß für die Interobserver-Variabilität bezeichnet, wobei hier kritisch anzumerken ist, dass bei der Berechnung der Bias zwischen den Untersuchern/Objekten nicht berücksichtigt ist, die Werte somit definitionsbedingt kleiner ausfallen als das oben angegebene Interobserver-Maß.

2.4.3 Sicher detektierbare Volumenänderungen

Für die klinische Praxis ist es wichtig zu wissen, ab welchem Betrag ein Unterschied zweier Messungen an einem Probanden (z. B. vor und nach Intervention) auf einer Volumenänderung beruht oder wegen Messfehler rein zufällig sein könnte.

Unter gewissen Voraussetzungen kann diese Frage mit einem einfachen Test beantwortet werden. Sind die Messwerte unabhängig und normalverteilt mit bekannter Varianz VAR(x), ist die Prüfgröße

$$Z = \frac{\left|x_1 - x_2\right|}{\sqrt{VAR(x_1 - x_2)}} = \frac{\left|x_1 - x_2\right|}{\sqrt{2 \cdot VAR(x)}} = \frac{\left|x_1 - x_2\right|}{\sqrt{2} \cdot \sigma}$$

standardnormalverteilt. Dann kann die Nullhypothese, es liegt keine Volumenänderung vor, mit einer Irrtumswahrscheinlichkeit von 0,05 abgelehnt werden, wenn Z > 1,96 ist (zweiseitiger Test).
Anders formuliert: die gemessenen Volumina unterscheiden sich mit einer Sicherheitswahrscheinlichkeit von 95%, wenn

$$\left|x_1 - x_2\right| > \Delta V \ \textit{mit}\ \Delta V = 1,96 \cdot \sqrt{2} \cdot \sigma = 2,77 \cdot \sigma$$

$$\text{d.h.}\ \left|x_1 - x_2\right| > 2,77 \cdot \text{Streuung (x)}$$

ausfällt.
Je nachdem, ob derselbe oder zwei verschiedene Untersucher die Messungen vorgenommen haben, muss dabei die Intra- oder die Interobserver-Streuung s_{intra} bzw. s_{inter} als Schätzung für σ eingesetzt werden.
Wenn die Berechnungen mit relativen Differenzen $d = \frac{x - v}{v}$ durchgeführt wurden, kann die mit 95% sicher detektierbare relative Volumenänderung zweier Messungen $\left|\frac{x_2 - x_1}{x_1}\right|$ (bezogen auf x_1) ähnlich wie oben bestimmt werden durch:

$$\left|\frac{x_2 - x_1}{x_1}\right| > \Delta V_{rel} = 1,96 \cdot \sqrt{2} \cdot Streuung\,(d)\,.$$

Dies ergibt sich aus folgenden Überlegungen:
Seien v_1 und v_2 die wahren Volumina der gemessenen Objekte und *VAR(d)* die Varianz der relativen Differenzen und weiterhin μ_i die Erwartungswerte von x_i, *VAR(x)* die Varianz und δ der Bias, d.h. $\mu_i = v_i + \delta$ (i = 1,2). Unter der Nullhypothese $v_1 = v_2 = v$, also $\mu_1 = \mu_2 = \mu$, hat die Zufallsvariable $\left|\frac{x_2 - x_1}{x_1}\right| = \left|\frac{x_2}{x_1} - 1\right|$ den Erwartungswert 0 und die Varianz

$$\frac{\mu_2^2}{\mu_1^2} \cdot \left(\frac{VAR(x_1)}{\mu_1^2} + \frac{VAR(x_2)}{\mu_2^2} \right) = \frac{2 \cdot VAR(x)}{\mu^2} = 2 \cdot VAR(\frac{x}{\mu})$$

(Voraussetzung: x_1 und x_2 sind statistisch unabhängig [Stuart 1994]).

Wenn der Bias δ klein gegenüber v ist, kann $VAR\left(\frac{x}{\mu}\right)$ durch $VAR\left(\frac{x}{v}\right) = VAR(d)$ ersetzt werden. Ist $\frac{x_2}{x_1}$ genügend genau normalverteilt, folgt für eine Sicherheitswahrscheinlichkeit von 95% die genannte Testvorschrift.

Allgemein lässt sich ΔV_β für beliebige Sicherheitswahrscheinlichkeiten β bei einem zweiseitigen Test und bekannter Streuung σ berechnen durch

$$\Delta V_\beta = C_\beta \cdot \sqrt{2} \cdot \sigma$$

mit $C_\beta = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$ (Φ^{-1}: inverse Standardnormalverteilung). Kann einseitig getestet werden, lautet die Formel $C_\beta = \Phi^{-1}(\beta)$.

Beispiele:

$\beta_{zweiseitig}$	$\beta_{einseitig}$	$C_\beta \cdot \sqrt{2} = \frac{\Delta V_\beta}{\sigma}$
50%	75%	0,95
60%	80%	1,19
70%	85%	1,47
80%	90%	1,81
90%	95%	2,33
95%	97,5%	2,77

Tab. 1: Sicherheitswahrscheinlichkeiten und Faktoren zur Berechnung von Volumenänderungen.

Wird die Formel nach β aufgelöst

$$\beta = 2 \cdot \Phi\left(\frac{\Delta V}{\sqrt{2} \cdot \sigma}\right) - 1,$$

kann zu einem gegebenen ΔV die zugehörige Sicherheitswahrscheinlichkeit bestimmt werden. Aus Fig. 1 kann β auch direkt abgelesen werden.

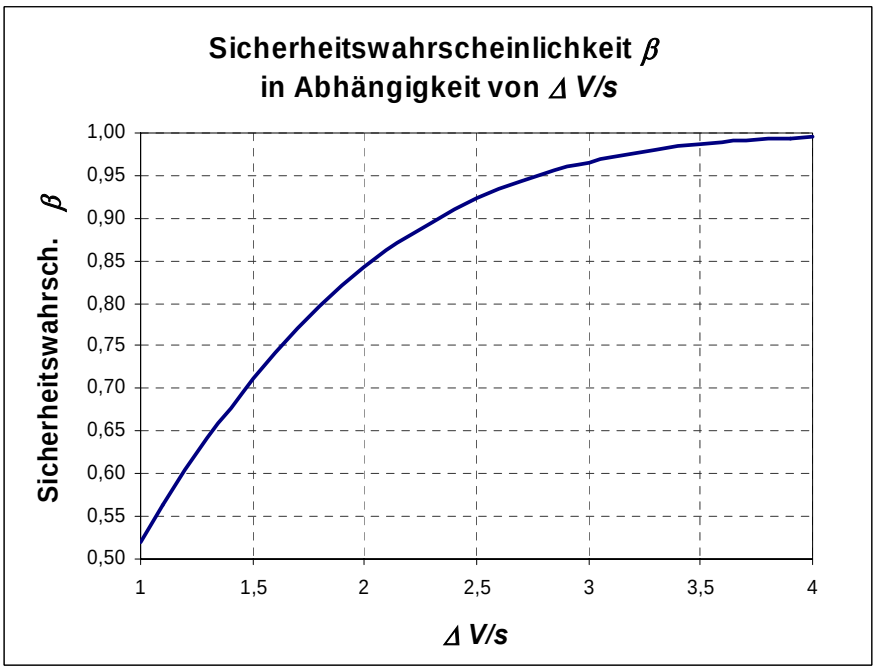


Fig. 1: Sicherheitswahrscheinlichkeit β in Abhängigkeit von Volumenänderung/Streuung bei zweiseitigem Test. Bei einem einseitigen Test ist $\beta_{einseitig} = (\beta_{zweiseitig} + 1)/2$.

Ein Beispiel:

Der oft benutzte Wert ΔV = 50% liefert bei einer Interobserver-Streuung von 30% eine Sicherheitswahrscheinlichkeit $\beta = 2 \cdot \Phi(1{,}18) - 1 = 0{,}76$, d.h. bei einer gemessenen Volumenänderung von 50% wird im Mittel in 24% aller Fälle eine Änderung angenommen, obwohl in Wirklichkeit keine vorliegt.

Gelegentlich wird als Schranke $\Delta V = s$ empfohlen, d.h. eine Änderung wird als solche akzeptiert, wenn sie den Wert der Streuung überschreitet. In diesem Fall ergibt sich $\beta = 2 \cdot \Phi(\frac{1}{\sqrt{2}}) - 1 = 0{,}52$, bei Messwertunterschieden an der Grenze *s* ist die Irrtumswahrscheinlichkeit also fast 50%.

2.4.4 Vergleich von Messverfahren: Die Methode nach Bland-Altman

Beim Vergleich von Messwerten muss beachtet werden, dass die Genauigkeit der Methoden von der Größe der Messwerte selbst abhängen kann. Unter Umständen müssen dann Richtigkeit und Präzision für verschiedene Messwertbereiche getrennt angegeben werden. Um solche Abhängigkeiten beurteilen zu können, kann die Methode von Bland-Altman [Bland 1986] herangezogen werden. Es handelt sich dabei um eine graphisch-statistische Methode, bei der die Differenzen gegen die Mittelwerte der Messwertpaare aufgetragen werden. Auf diese Weise werden eventuelle Abhängigkeiten sofort sichtbar. Bland und Altman empfehlen, den Mittelwert der Differenzen als Maß für die Richtigkeit und die 95%-Streubereiche (Limits of Agreement) als Maß für die Präzision einzuzeichnen. Bei starker Größenabhängigkeit der Genauigkeit werden Formeln für nicht-parallele Streubereiche angegeben. Wächst die Streuung der Werte mit der Größe, kann versucht werden, anstatt der absoluten Differenzen die Verhältnisse der Wertepaare in Ordinatenrichtung aufzutragen, oder – äquivalent – zu einer logarithmischen y-Skala überzugehen, um so eventuell eine über den ganzen relevanten Messbereich gültige Präzision angeben zu können.

Bland und Altman empfehlen weiterhin, in Abszissenrichtung die Mittelwerte aufzutragen, selbst wenn gegen einen Goldstandard (bzw. exakte Messwerte) verglichen wird, weil die Differenzen gegen einen der Werte immer (üblicherweise negativ) korreliert sind [Bland 1995]. Bei hohen Korrelationen zwischen den Messwerten, was eigentlich immer der Fall sein sollte, ist die Größenordnung aber vernachlässigbar.

Da bei den Untersuchungen der vorliegenden Studie zur Intra- und Interobserver-Variabilität zu mehreren 2D-Messwerten *(V_{2D})* nur ein 3D-Wert als Referenzvolumen *(V_{ref})* vorliegt und dadurch eine strenge Abhängigkeit der Differenzen bzw. Verhältnisse zu den Mittelwerten vorliegt, werden die (relativen) Differenzen *(D)* immer gegen die 3D-Volumina aufgetragen und nach folgender Formel berechnet:

$$D = \frac{V_{2D} - V_{ref}}{V_{ref}}$$

In verschiedenen Arbeiten beschreiben Bland und Altman die Vorzüge ihrer Methode gegenüber Korrelations- und Regressionsverfahren [Bland 1986, Bland 1999].

2.4.5 Intra- und Interobserver-Reliabilität: der multivariate Ansatz von Eliasziw et al.

Ein in der Literatur oft anzutreffendes Verfahren, um die Zuverlässigkeit von Messungen zu beurteilen, ist das von Fleiss und Shrout [Fleiss 1978] entwickelte Konzept der Intra-Klassen-Korrelation (ICC). Es beruht auf einem varianzanalytischen Modell und kann als Maß der Varianzaufklärung zwischen den Beurteilern aufgefasst werden. Der ICC-Koeffizient wird wie ein Korrelationskoeffizient interpretiert: ICC = 0 bedeutet, dass keine und ICC = 1, dass eine strenge Übereinstimmung der Beurteiler vorliegt. Je nachdem, ob die Untersucher als zufällige oder als feste Faktoren verstanden werden, resultieren verschiedene Vorschriften zur Berechnung des ICC.

Wenn jedes Objekt mehr als einmal vom gleichen Untersucher gemessen wird, ist dieses Verfahren nicht mehr direkt anwendbar. Für diesen Fall geben Eliasziw et al. [Eliasziw 1994] ein Verfahren an, das als Verallgemeinerung des ICC aufgefasst werden kann: mittels einer zweifaktoriellen Varianzanalyse werden sog. Intra- und Interobserver-Reliabilitätskoeffizienten bestimmt. Dabei gibt der Intraobserver-Reliabilitätskoeffizient an, wie konsistent und reproduzierbar die Messungen sind, der Interobserver-Reliabilitätskoeffizient bestimmt die Konsistenz der Messungen mehrerer Untersucher an einem Objekt und gibt an, inwieweit die Untersucher austauschbar sind.
Die Werte der Reliabilitätskoeffizienten hängen stark vom Verhältnis der Objekt-Varianzen zu den Observer-Varianzen ab; ist die Objekt-Variabilität gering, resultieren kleine Reliabilitätskoeffizienten. Die Koeffizienten zweier Messverfahren sind daher nur miteinander vergleichbar, wenn die Objektgrößen in der Stichprobe etwa gleich verteilt sind.

2.4.5.1 Das Modell

Zur Berechnung der Varianzen wird eine zweifaktorielle Varianzanalyse (ANOVA) durchgeführt, wobei eine eventuelle Interaktion zwischen Untersucher und Objekt berücksichtigt wird (saturiertes Modell):

$$x_{ijk} = \mu + R_j + S_i + (RS)_{ij} + \varepsilon_{ijk}$$

Dabei ist x_{ijk} die *k*-te Messung des Untersuchers *j* am Probanden *i*; μ ist der Gesamterwartungswert, R_j der Effekt des *j*-ten Untersuchers, S_i der Effekt des *i*-ten Probanden. $(RS)_{ij}$ und ε_{ijk} repräsentieren den Inter- und Intraobserver-Fehler.
Der Nomenklatur von Eliasziw et al. [Eliasziw 1994] folgend, werden die Varianzen der Messungen des Untersuchers *(R)*, des Objekts *(S)*, des Interaktionsterms *(RS)* und einer nicht weiter erklärbaren Restvarianz jeweils bezeichnet als σ_R^2, σ_S^2, σ_{RS}^2 und σ_{err}^2. Die Berechnung der einzelnen Varianzen hängt davon ab, ob der Faktor Untersucher als fest oder zufällig aufgefasst wird. Hier und des Weiteren wird angenommen, dass die Untersucher eine gewisse Grundgesamtheit repräsentieren (zufälliger Faktor). Die für eine ANOVA übli-

chen Annahmen müssen erfüllt sein, d.h. im Falle zufälliger Untersucher sollten R_j, S_i, $(RS)_{ij}$ und ε_{ijk} paarweise unabhängig normalverteilt sein mit Erwartungswert 0 und den jeweiligen Varianzen. Die Gesamtvarianz ergibt sich als Summe der Einzelvarianzen:

$$\sigma^2_{total} = \sigma^2_R + \sigma^2_S + \sigma^2_{RS} + \sigma^2_{err}$$

Üblicherweise geben die Statistikprogramme nicht die Schätzer der gesuchten Varianzen an, sondern nur die jeweiligen mittleren Quadratsummen (Mean Squares = *MS*). Aus diesen lassen sich die Werte unter den gegebenen Annahmen berechnen durch:

$$\hat{\sigma}^2_R = \frac{(MS_R - MS_{RS})}{m \cdot n}$$

$$\hat{\sigma}^2_S = \frac{(MS_S - MS_{RS})}{m \cdot r}$$

$$\hat{\sigma}^2_{RS} = \frac{(MS_{RS} - MS_{err})}{m}$$

$$\hat{\sigma}^2_{err} = MS_{err}$$

wobei im Fall der Probanden Studie m = Anzahl der Messungen, r = Anzahl der Untersucher und n = Anzahl der Objekte ist.

2.4.5.2 Intra- und Interobserver-Fehler SEM_{intra} und SEM_{inter}

Durch die Einzelvarianzen, die durch den Untersucher (σ_R^2), das untersuchte Objekt (σ_S^2), die Interaktion zwischen Untersucher und Objekt (σ_{RS}^2) und die Messung selbst (σ_{err}^2) in die Gesamtvarianz einfließen, können die Intra- und Interobserver-Streuungen oder Standardmessfehler SEM_{intra} und SEM_{inter} (Standard Error of Measurement) bestimmt werden:

$$SEM_{intra} = \sqrt{\sigma_{err}^2}$$

$$SEM_{inter} = \sqrt{\sigma_R^2 + \sigma_{RS}^2 + \sigma_{err}^2}$$

2.4.5.3 Intra- und Interobserver-Reliabilitätskoeffizienten ρ_{intra} und ρ_{inter}

Die Intra- und Interobserver-Reliabilitätskoeffizienten sind definiert als:

$$\rho_{intra} = \frac{1 - SEM_{intra}^2}{\sigma_{total}^2} = \frac{(\sigma_R^2 + \sigma_S^2 + \sigma_{RS}^2)}{\sigma_{total}^2}$$

$$\rho_{inter} = \frac{1 - SEM_{inter}^2}{\sigma_{total}^2} = \frac{\sigma_S^2}{\sigma_{total}^2}$$

Landis und Koch [Landis 1977] geben Bereiche der **Reliabilitätskoeffizienten** an, die sie wie folgt interpretieren:

Intervall	Bewertung
0,0 – 0,20	gering („slight“)
0,21 – 0,40	ausreichend („fair“)
0,41 – 0,60	mäßig („moderate“)
0,61 – 0,80	beträchtlich („substantial“)
0,81 – 1,00	fast perfekt („almost perfect“)

Tab.2: Einteilung der Reliabilitätskoeffizienten nach Landis und Koch.

2.4.5.4 Absolute und relative Abweichungen

Bei der Beurteilung dieser Koeffizienten muss beachtet werden, dass sie nicht nur von den Messfehlern, sondern auch von der Größenverteilung der untersuchten Objekte abhängen. Unterscheiden sich die wahren Werte im Vergleich zu den Messfehlern nur wenig, ergeben sich niedrige Reliabilitätskoeffizienten [Wirtz 2002]. Wenn andererseits, wie im vorliegenden Fall, der absolute Messfehler mit der Größe zunimmt, sollten die Größenunterschiede der Objekte nicht zu extrem ausfallen, weil dann die Voraussetzungen der Varianzanalyse (gleiche Streuungen) nicht erfüllt sind und die Ergebnisse unbrauchbar werden. Ein Beispiel: Wenn sich bei einer Gruppe mit einem mittleren Schilddrüsenvolumen von 18 ml ein minimal detektierbarer Volumenunterschied von 6 ml ergibt, so würde das bei einem Probanden mit einem Lappenvolumen von 5 ml bedeuten, dass erst ein Zuwachs um mehr als das doppelte sicher festgestellt werden könnte.

Aus diesem Grund wurde das Verfahren nochmals auf eine "Normalgruppe" angewandt, bei der Probanden mit sehr großen und sehr kleinen Schilddrüsenvolumina ausgeschlossen wurden. Um den Einfluss der Schilddrüsengrößen auszuschalten, wurden die Berechnungen außerdem, einem Vorschlag von Tong et al. [Tong 1998] folgend, mit normierten Werten $\frac{V_{2D} - V_{ref}}{V_{ref}}$ durchgeführt.

2.4.6 Auswertesoftware

Für die Datenverarbeitung wurden die SPSS 12.0 Standardsoftware (SPSS Inc., Chicago, IL) und das Programm Excel (Microsoft Corp., Seattle, WA) verwendet.

3. ERGEBNISSE

Im folgenden Teil der Arbeit werden zunächst die Ergebnisse der Phantom-Studie, dann diejenigen der Probanden-Studie aufgeführt und graphisch dargestellt. Bei der Probanden-Studie wurde zusätzlich eine Varianzanalyse durchgeführt, deren Ergebnisse tabellarisch zusammengefasst und durch ein Beispiel erläutert werden.

3.1 Phantom-Studie

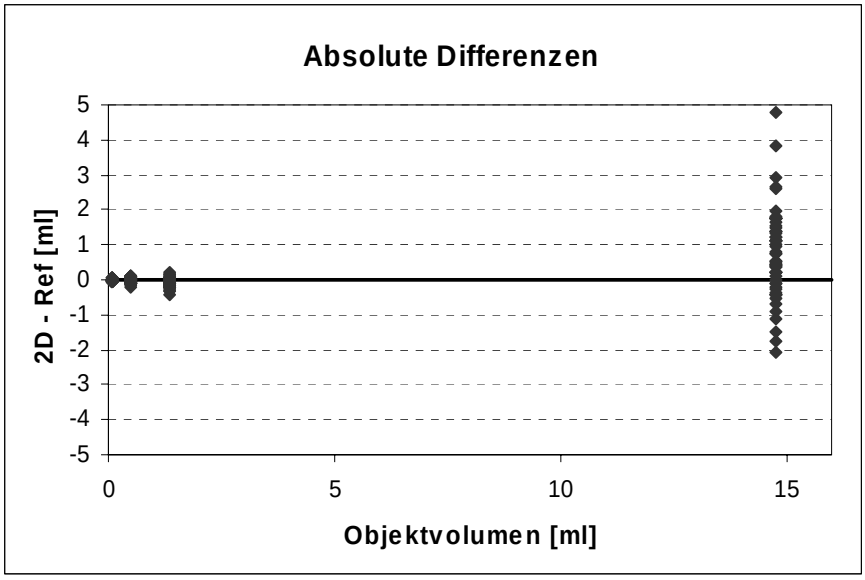
3.1.1 Referenzwerte

Objekt	Volumen [ml]
Knoten klein	0,1
Knoten mittel	0,51
Knoten groß	1,37
Lappen	14,75

Tab. 3: Exakte Referenzwerte der Phantom-Objekte in ml (identische Volumina jeweils für die Objekte rechts und links).

In Tabelle 3 sind die exakten Volumina der Phantomobjekte aufgeführt. Der große Phantom-Knoten ist 13,7 mal größer als der kleinste und um einen Faktor von ca. 11 kleiner als die Lappen. Dies muss bei der Beurteilung der Messgenauigkeiten berücksichtigt werden.

3.1.2 Absolute und relative Differenzen



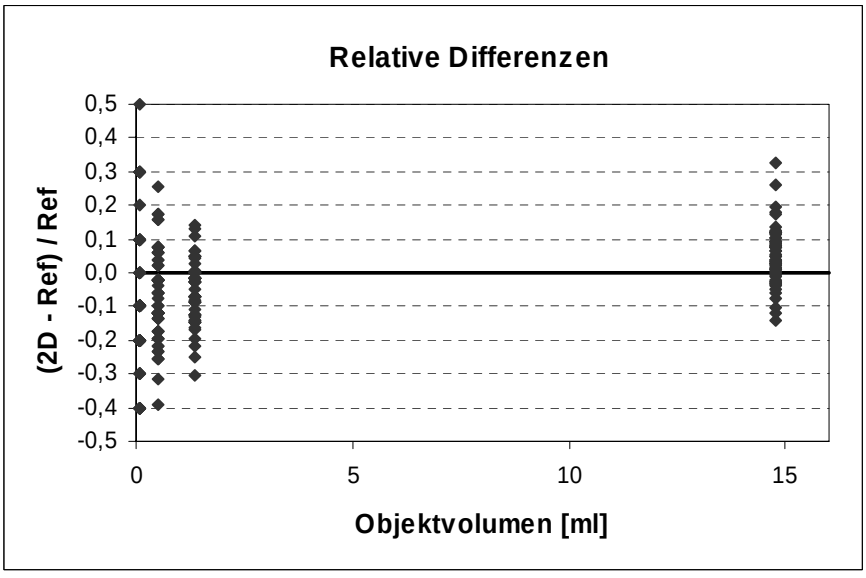


Fig. 2a, b. Absolute und relative Differenzen in Abhängigkeit von den exakten Objektvolumina.

In Fig. 2a sind, in Anlehnung an die Methode nach Bland-Altman [Bland 1986], alle 270 Messwerte als Abweichungen von den wahren Objektvolumina gegen die exakten Größen eingetragen. Zu erkennen sind die Größenverhältnisse der Objekte und eine deutliche Zunahme des Messfehlers (Streuung) mit deren Vo-

lumen. Wegen der großen Volumenunterschiede werden bei den untersucher- und objektabhängigen Fehlerbetrachtungen Knoten und Lappen getrennt behandelt. Um die Intraobserver-Variabilität (Messungen eines Untersuchers von allen Objekten) beurteilen zu können, sollten die Messvarianzen der Objekte vergleichbar sein. Daher wird im Folgenden mit auf das jeweilige Referenzvolumen bezogenen Differenzen gerechnet (relative Differenz = $\frac{V_{2D} - V_{ref}}{V_{ref}}$).

Fig. 2b zeigt die relativen Differenzen in Abhängigkeit von den wahren Objektvolumina. Die Streubreite der Messfehler wird durch die Transformation vom Volumen weitgehend unabhängig. Die kleinen Knoten werden mit relativ geringerer Auflösung gemessen und haben jetzt die größte Fehler-Varianz. Abzulesen ist außerdem, dass die Knoten offenbar unter- und die Lappen überschätzt werden.

Bei den Interobserver-Variabilitäten (Messungen eines Objektes von allen Untersuchern) erhält man die absoluten Differenzen durch Multiplikation mit dem jeweiligen exakten Volumen.

3.1.3 Übersicht über Messergebnisse und Methodenparameter

Mittelwerte und Streuungen, relative Differenzen										
Unter-sucher	Knoten							Lappen		
	klein		mittel		groß		alle	links	rechts	beide
	-10 dB	-5 dB	-10 dB	-5 dB	-10 dB	-5 dB				
1	-0,2333 0,1528	0,2000 0,1732	-0,1503 0,0742	0,0196 0,0679	-0,0438 0,1175	0,0414 0,0694	-0,0277 0,1738	0,0642 0,0490	0,0689 0,0437	0,0666 0,0416
2	-0,1000 0,0000	-0,1333 0,1155	-0,1242 0,0113	0,0915 0,0792	-0,0195 0,0084	-0,0462 0,0464	-0,0553 0,0943	0,0025 0,0192	0,0682 0,0157	0,0354 0,0393
3	-0,1667 0,0577	-0,3000 0,1000	-0,2288 0,1006	-0,1895 0,0113	-0,0827 0,0548	-0,1800 0,0497	-0,1913 0,0892	0,0405 0,0772	0,0409 0,0691	0,0407 0,0655
4	0,0333 0,1155	-0,0333 0,1155	-0,0523 0,0566	0,0392 0,0519	-0,0024 0,0694	-0,0584 0,0386	-0,0123 0,0785	0,0409 0,0517	0,0610 0,0801	0,0510 0,0613
5	0,1000 0,1000	-0,2667 0,1528	-0,1111 0,0816	0,0980 0,0707	0,0024 0,0513	-0,0852 0,0621	-0,0437 0,1533	0,0933 0,0901	0,1073 0,0823	0,1003 0,0776
6	-0,2000 0,0000	-0,0667 0,1155	-0,1176 0,0588	0,0588 0,1709	-0,0925 0,1056	-0,0097 0,0084	-0,0713 0,1174	0,0364 0,0721	0,0768 0,0442	0,0566 0,0579
7	0,0000 0,2646	0,1333 0,3215	-0,1046 0,0226	-0,1373 0,1556	-0,1071 0,1729	0,0389 0,1064	-0,0294 0,1947	-0,0809 0,0825	0,1261 0,1228	0,0226 0,1470
8	-0,1667 0,0577	0,1000 0,2000	-0,2026 0,0599	-0,0588 0,0392	-0,1314 0,0126	-0,1484 0,0901	-0,1013 0,1314	0,0373 0,0360	0,0258 0,0808	0,0315 0,0563
9	-0,3000 0,1000	-0,3333 0,1155	-0,0654 0,2012	-0,1961 0,2066	0,0292 0,0379	-0,0049 0,1388	-0,1451 0,1894	-0,0362 0,0349	0,1846 0,1365	0,0742 0,1502
alle Unter-sucher	-0,1148 0,1634	-0,0778 0,2359	-0,1285 0,0936	-0,0305 0,1478	-0,0497 0,0897	-0,0503 0,0982	-0,0753 0,1496	0,0220 0,0721	0,0844 0,0834	0,0532 0,0834

Tab. 4: Ergebnisse der Messungen: Mittelwerte (oben) und Streuungen (unten) der Messungen pro Untersucher und Objekt, relative Differenzen. In der letzten Spalte sind die Mittelwerte und Streuungen der Messungen pro Untersucher über alle Objekte eingetragen, sie repräsentieren den Intraobserver-Bias und die Intraobserver-Variabilität. In der letzten Zeile: Bias und Variabilität der Werte aller Untersucher pro Objekt (Interobserver).

In Tab. 4 sind die Mittelwerte (als Maß für die Richtigkeit) und Streuungen (als Maß für die Präzision) der relativen Messdifferenzen pro Untersucher und Objekt zusammengefasst. Jede Zelle repräsentiert also 3 Messungen. Die absoluten Abweichungen pro Objekt erhält man durch Multiplikation mit dem jeweiligen Referenzvolumen. Anhand der Werte in der Zeile "alle Untersucher" lassen sich die Interobserver-Variabilitäten beurteilen. Auf die Intra- und Interobserver-Variabilität wird weiter unten ausführlicher eingegangen.

Minima und Maxima rel. Differenzen		Knoten							Lappen		
		klein		mittel		groß		alle	links	rechts	beide
		-10 dB	-5 dB	-10 dB	-5 dB	-10 dB	-5 dB				
MW	Min.	-0,3000	-0,3333	-0,2288	-0,1961	-0,1314	-0,1800	-0,1913	-0,0809	0,0258	0,0226
	Max.	0,1000	0,2000	-0,0523	0,0980	0,0292	0,0414	-0,0123	0,0933	0,1846	0,1003
	Mittelwert	-0,1148	-0,0778	-0,1285	-0,0305	-0,0497	-0,0503	0,0600	0,0220	0,0844	0,0243
Streu-ung	Min.	0,0000	0,1000	0,0113	0,0113	0,0084	0,0084	0,0785	0,0192	0,0157	0,0393
	Max.	0,2646	0,3215	0,2012	0,2066	0,1729	0,1388	0,1947	0,0901	0,1365	0,1502
	Mittelwert	0,0943	0,1566	0,0741	0,0948	0,0700	0,0678	0,1358	0,0570	0,0750	0,0774

Tab. 5. Spannweiten und Mittelwerte der relativen Abweichungen aller Untersucher pro Objekt. Der Mittelwert der Streuungen beschreibt den reinen mittleren Messfehler. Multiplikation mit dem jeweiligen Referenzvolumen ergibt die Werte für die absoluten Differenzen.

Tab. 5 gibt einen Überblick über die Maxima und Minima der Mittelwerte und Streuungen aus Tab. 4, außerdem sind die mittleren Streuungen angegeben; sie repräsentieren den reinen mittleren Messfehler der Untersucher pro Objekt.

relative Differenzen	Seite	Mittel-wert	s_{inter}	95%-Streuber. (Limits of Agreement)		Std-Fehler	95%-Konf.-Intervall		s_{intra}
Knoten klein	- 10 dB	-0,1148	0,1634	-0,4351	0,2054	0,0314	-0,1764	-0,0532	0,1217
	- 5 dB	-0,0778	0,2359	-0,5401	0,3846	0,0454	-0,1668	0,0112	0,1699
	alle	-0,0963	0,2018	-0,4919	0,2993	0,0275	-0,1501	-0,0425	0,1805
Knoten mittel	- 10 dB	-0,1285	0,0936	-0,3120	0,0549	0,0180	-0,1639	-0,0932	0,0905
	- 5 dB	-0,0305	0,1478	-0,3201	0,2591	0,0284	-0,0862	0,0252	0,1137
	alle	-0,0795	0,1321	-0,3385	0,1795	0,0180	-0,1148	-0,0443	0,1228
Knoten groß	- 10 dB	-0,0497	0,0897	-0,2255	0,1260	0,0173	-0,0836	-0,0159	0,0863
	- 5 dB	-0,0503	0,0982	-0,2427	0,1421	0,0189	-0,0873	-0,0133	0,0772
	alle	-0,0500	0,0931	-0,2325	0,1325	0,0127	-0,0748	-0,0252	0,0854
Knoten	- 10 dB	-0,0977	0,1320	-0,3564	0,1610	0,0016	-0,1009	-0,0945	0,1149
	- 5 dB	-0,0529	0,1309	-0,3095	0,2037	0,0016	-0,0561	-0,0497	0,1506
	alle	-0,0753	0,1496	-0,3685	0,2179	0,0118	-0,0983	-0,0522	0,1421
Lappen	- 10 dB	0,0220	0,0721	-0,1193	0,1633	0,0139	-0,0052	0,0492	0,0616
	- 5 dB	0,0844	0,0834	-0,0790	0,2479	0,0161	0,0530	0,1159	0,0832
	alle	0,0532	0,0834	-0,1103	0,2167	0,0113	0,0310	0,0755	0,0869

Tab. 6: Mittelwerte mit Intra- und Interobserver-Streuungen (s) sowie den 95%-Streubereichen (Limits of Agreement) und 95%-Konfidenzintervallen für die Mittelwerte. Relative Differenzen.

s_{err}						
Knoten						Lappen
klein	mittel	groß	-10 dB	-5 dB	alle	beide
0,1478	0,1027	0,0819	0,1007	0,1262	0,1141	0,0732

Tab. 7. Methodenimmanenter Messfehler s_{err}.

Aus Tab. 6 lassen sich die Größenordungen der Interobserver-Bias und -Streuungen der Messungen ablesen, wobei im Hinblick auf die Beurteilung der Echogenität die Werte für die -5 dB und -10 dB Objekte jeweils auch insgesamt aufgelistet sind. Zusätzlich sind die 95%-Streubereiche (Limits of Agreement), die Standardfehler und deren 95%-Konfidenzintervalle aufgelistet. Enthält das jeweilige Konfidenzintervall nicht den Wert Null, ist die mittlere Abweichung signifikant.

Werden jeweils alle Knoten der Echogenität -5 dB und -10 dB zusammengefasst, ergeben sich als relative Abweichungen -5,3% bzw. -9,8%. Für die Messungen der Lappen (gleiche Echogenität) liegt die relative Abweichung bei +5,3%. Es ist zu erkennen, dass weder die Unterschätzung der Knotenvolumina noch die Überschätzung der Lappen zufällig ist, der jeweilige Bias also als signifikant gelten kann.

Aus den Streuungen s_{inter} und s_{intra} können durch Multiplikation mit $1{,}96 \cdot \sqrt{2}$ (2,77) die sicher detektierbaren Volumenänderungen berechnet werden. Als Interobserver-Werte ergeben sich für die kleinen Knoten 56%, für die mittleren Knoten 37%, für die großen Knoten 26% und für die Lappen 23%; für alle Knoten liegt die sicher detektierbare Volumenänderung bei 41%. Die Intraobserver-Werte sind geringfügig kleiner.

Aus Tab. 7 lässt sich der Fehler der Messmethode s_{err} ersehen, der $\hat{\sigma}_{err}$ entspricht. Für die Knoten liegt er bei 11%; werden die Lappen gemessen, beträgt er 7,3%.

Vergleicht man die relativen Abweichungen der Knotenmessungen verschiedener Echogenität (-5 dB und -10 dB) miteinander, so ergibt ein zweiseitiger gepaarter t-Test eine Irrtumswahrscheinlichkeit von 5,2%, was nur knapp über dem üblichen Signifikanzniveau von 5% liegt. Dies deutet darauf hin, dass die Echogenität einen Einfluss auf die Messungen haben könnte. Ein F-Test auf Unterschiede zwischen -10 dB- und -5 dB-Messvarianzen (Zähler- und Nenner-Freiheitsgrade: $(m-1)\cdot s\cdot r = 54$) ergibt $F = \frac{s^2_{-5\,dB}}{s^2_{-10\,dB}} = 1{,}57$, was ebenfalls nur knapp unter der Testschranke (1,60 = $F_{54,54,95\%}$) liegt.

3.1.4 Intraobserver-Variabilität

3.1.4.1 Intraobserver-Variabilität der Knoten

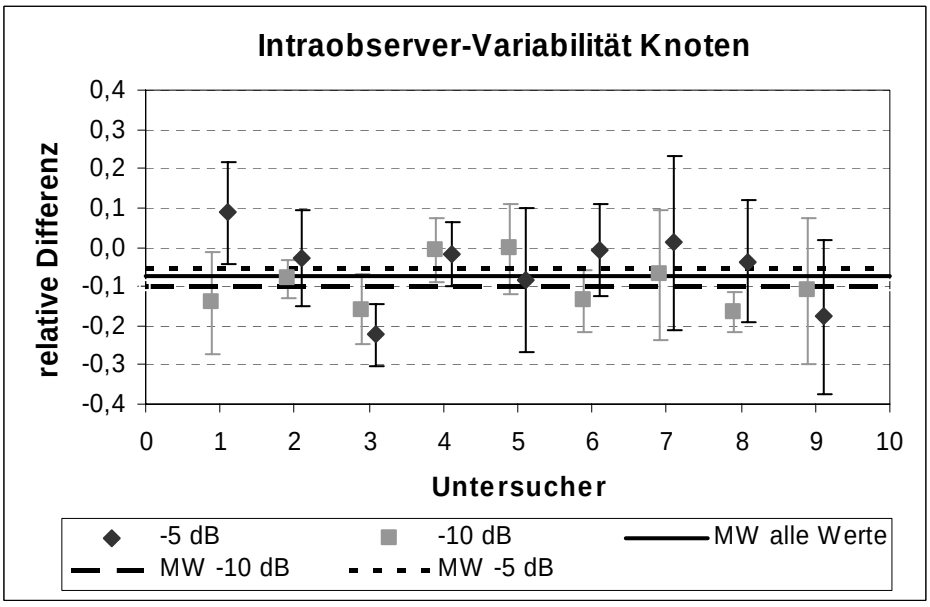


Fig. 3*: Intraobserver-Variabilität (Mittelwerte und Streuungen) aller Knoten, differenziert nach Echogenität (Rauten: -5 dB-Knoten, Quadrate: -10 dB-Knoten). Die horizontale durchgezogene Linie stellt den Mittelwert aller Knotenmessungen dar; die horizontale gepunktete Linie steht für den Mittelwert der -5 dB-Knoten, die horizontale gestrichelte Linie für den Mittelwert der -10 dB-Knoten.*

Um die Intraobserver-Variabilität bei den Knotenmessungen zu veranschaulichen, werden für jeden Untersucher (1 bis 9) jeweils getrennt für die Echogenitäten -5 dB und -10 dB die Mittelwerte der standardisierten Differenzen aufgetragen. Diese erstrecken sich von -22,3% bis +8,7% (Median -5,3%). Wegen der großen Volumenunterschiede werden Knoten und Lappen in zwei verschiedenen Diagrammen dargestellt (vgl. auch Fig. 4). Ein Datenpunkt (Raute oder Quadrat) enthält den Mittelwert der Messungen aller isoechogener Knoten pro Untersucher. Die Fehlerbalken entsprechen den Standardabweichungen als ein Maß für die Präzision.

Aus Fig. 3 lässt sich weiterhin ablesen, dass der jeweilige Bias (-9,8% für die -5 dB-Knoten und -5,3% für die -10 dB-Knoten) im Vergleich zu den Streuungen relativ gering ist. Zusätzlich fällt auf, dass bezüglich der Richtigkeit die Mittelwerte mancher Untersucher (z.B. 1 und 8) stark zwischen rechtem und linkem Lappen differieren, während andere Untersucher (z.B. 2 und 4) sehr konsistent in ihren Messungen sind. Die Messungen einiger Untersucher (z.B. 7 und 9) weisen zudem eine große Streubreite auf, sind also sehr unpräzise. Bei Untersucher 4 dagegen vereinigt sich eine hohe Konsistenz der Messungen mit einer geringen Streubreite.

3.1.4.2 Intraobserver-Variabilität der Lappen

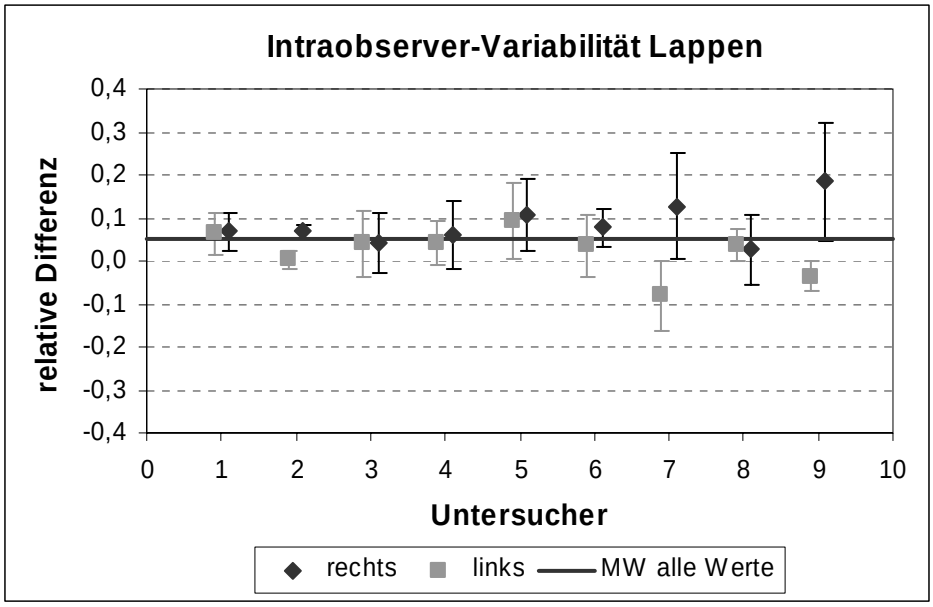


Fig. 4: Intraobserver-Variabilität (Mittelwerte und Streuungen) der Lappen. Anders als bei den Knoten sind beide Lappen isoechogen. Die horizontale durchgezogene Linie stellt den Mittelwert aller Lappenmessungen dar.

In Fig. 4 ist die Intraobserver-Variabilität der Lappen dargestellt. Ein Datenpunkt (Raute oder Quadrat) enthält den Mittelwert der Messungen eines Lappens pro Untersucher. Die Fehlerbalken stellen die zugehörige Streubreite dar. Für den einzelnen Untersucher erstrecken sich die Mittelwerte der standardisierten Differenzen von -8,1% bis +18,4% (Median 5,3%). Die horizontale durchgezogene Linie steht für den Mittelwert über alle Lappenmessungen und zeigt bezüglich der Richtigkeit eine systematische, statistisch signifikante *Über*schätzung der Lappenvolumina im 2D-Ultraschall gegenüber den Referenzvolumina ($p < 0,05$).

Anders als bei den Knoten weisen beide Lappen die gleiche Echogenität auf. Daher ist erwartungsgemäß eine größere Konsistenz der Messungen festzustellen als bei den Knoten. Umso erstaunlicher ist jedoch, dass bei den Untersuchern 2,7 und 9 eine große Differenz der Mittelwerte zwischen rechtem und linkem Lappen besteht. Zusätzlich sind die Untersucher 7 und 9 sehr unpräzise

bei ihren Messungen. Im Gegensatz zu den Knotenmessungen liegt der Bias der meisten Untersucher in der Größenordung der Streuungen.

3.1.5 Interobserver-Variabilität der Phantom-Objekte

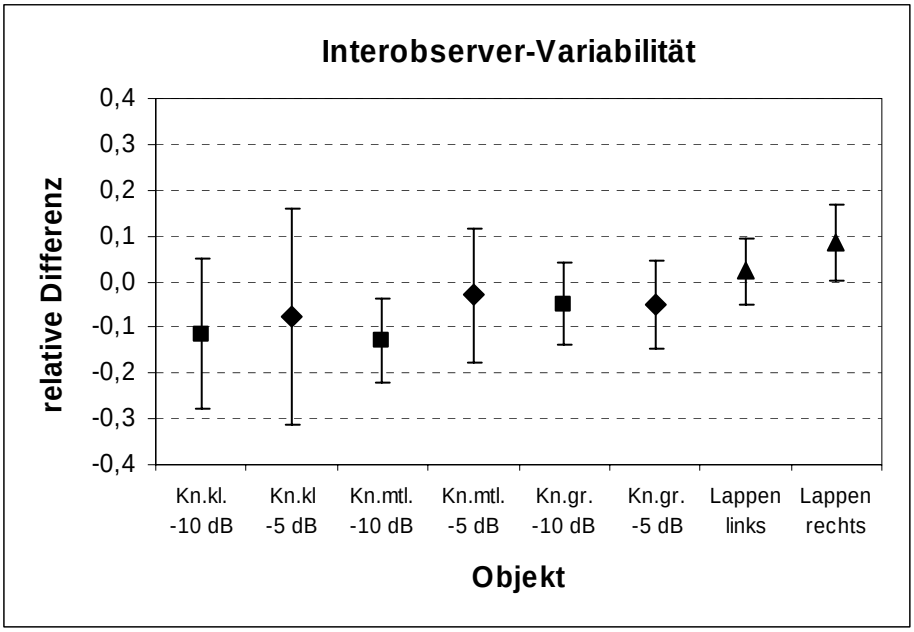


Fig. 5: Interobserver-Variabilität: Mittelwerte und Streuungen aller Phantomobjekte, differenziert nach Objektgröße (von links nach rechts zunehmend) und für die Knoten zusätzlich nach Echogenität. Bei den isoechogenen Lappen wird zwischen rechts und links unterschieden.

In Fig. 5 sind noch einmal die Daten für die Interobserver-Variabilität der einzelnen Phantomobjekte graphisch dargestellt. Dazu werden für die Messungen der Knoten und Lappen die Mittelwerte der standardisierten Differenzen aufgetragen und nach Objektgröße und für die Knoten zusätzlich nach Echogenität differenziert. Bei den isoechogenen Lappen wird zwischen rechts und links unterschieden. Von links nach rechts nimmt das Objektvolumen zu (0,1 ml bis 14,75 ml). Quadrate symbolisieren die -10 dB-Knoten, Rauten die -5 dB-Knoten, Dreiecke stehen für die Lappen. Die Fehlerbalken zeigen die Standardabweichungen an (jeweils drei Messungen für alle neun Untersucher).

Aus dieser Synopsis der Phantom-Messungen wird nochmals deutlich, dass mit zunehmender Objektgröße der Bias ansteigt, die Streubreite der relativen Messungen jedoch abnimmt.

3.1.6 Zufälliger Untersucher-Fehler (Random Observer Error)

Während bei der Intraobserver-Variabilität als Maß für die Präzision die Streuung der Messungen pro Untersucher über die Messungen aller Objekte angegeben wurde, führen Tong et al. [Tong 1998] den zufälligen Untersucherfehler (Random Observer Error) als die mittlere Streuung der Mehrfachmessungen der Objekte ein, bezogen auf einen Untersucher. Der zufällige Untersucherfehler ist also die mittlere reine Mess-Streuung pro Untersucher; der Bias bleibt dabei unberücksichtigt.

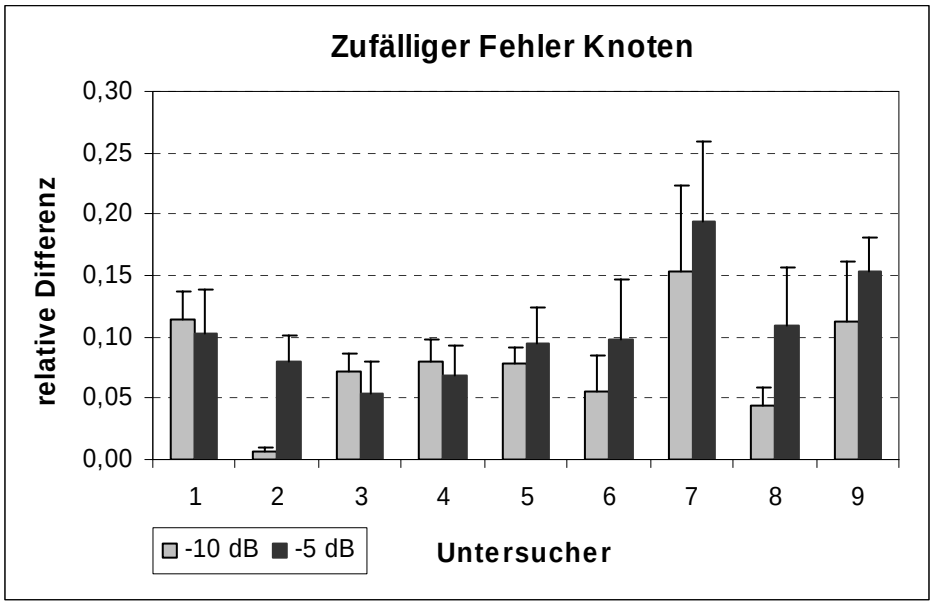


Fig. 6: Zufälliger Fehler nach Tong et al. (Random Observer Error) der Knoten, differenziert nach Echogenität: Der Mittelwert der Streuungen der Messwerte eines Untersuchers pro Objekt. Die Fehlerbalken entsprechen den Standardfehlern der jeweiligen Mittelwerte der Streuungen.

Fig. 6 visualisiert die Messgenauigkeiten der verschiedenen Knoten-Echogenitäten, d.h. alle Knotenmessungen gleicher Echogenität sind zusammengefasst. Obwohl z.T. deutlich unterschiedlich, ist kein eindeutiger Trend zu erkennen, d.h. aus den Daten kann nicht abgeleitet werden, dass etwa Objekte mit größerer Echogenität genauer gemessen werden.

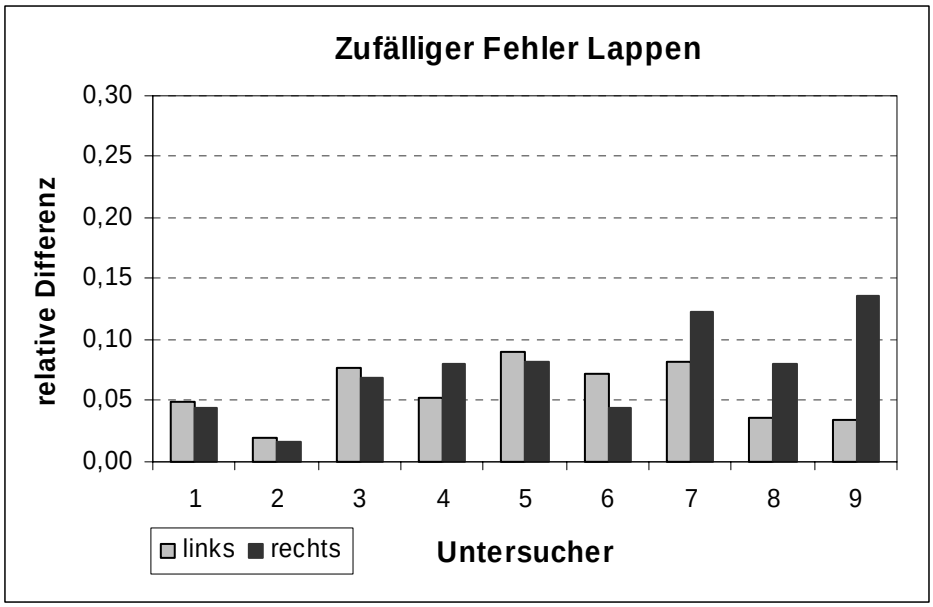


Fig. 7: Zufälliger Fehler (Random Observer Error) der Lappen, differenziert nach Lokalisation. Ein eindeutiger Unterschied zwischen dem rechten und dem linken Lappen lässt sich nicht erkennen. Die Messfehler sind somit vergleichbar.

Der Vollständigkeit halber sind in Fig. 7 die mittleren Messfehler der Lappen, getrennt nach Lokalisation, graphisch dargestellt. Jeder Balken repräsentiert hier die Streuung der Messungen eines Objektes (des Lappens); daher kann kein Standardfehler eingezeichnet werden. Im Vergleich mit den Ergebnissen der Knoten ergeben sich ähnliche Größenordungen der relativen Differenzen. Bei gleicher Echogenität der Lappen lassen sich zwar individuelle Schwankungen erkennen, eine Präferenz für eine Lokalisation lässt sich jedoch nicht ableiten. Offenbar zeigten die Untersucher keine Tendenz, eine Seite genauer zu messen.

3.1.7 Zusammenfassung der Phantom-Ergebnisse

Die folgenden Resultate ergeben sich zusammenfassend aus den Messungen der 8 Phantom-Objekte durch 9 Untersucher mittels des 2D-Ultraschall-Verfahrens. Die im Methodenteil beschriebene multivariate Reliabilitätsanalyse nach Eliasziw et al. [Eliasziw 1994] wurde bei dieser Studie nicht durchgeführt,

weil die Objekte nicht zufällig ausgewählt waren, sondern ausgesucht wurden, um Messungen an Objekten verschiedener Größenklassen zu untersuchen.

1. Die Messungen der Knoten- *und* Lappenvolumina weisen einen statistisch signifikanten Bias auf.
2. Knotenvolumina werden signifikant unterschätzt, und zwar um 9,6% für kleine bis 5,0% für große Knoten. Dabei nehmen die Interobserver-(Intraobserver-)Streuungen s_{inter} bzw. s_{intra} von 20% (18%) auf 9,3% (8,5%) ab.
3. Die Echogenität scheint einen, wenn auch geringen Einfluss auf die Messgenauigkeit der Knoten zu haben.
4. Der Bias bezogen auf alle Knoten beträgt -7,5%, die mittleren Streuungen s_{inter} bzw. s_{intra} liegen bei 15% (14%).
5. Die Lappenvolumina werden im Durchschnitt um 5,0% von den Untersuchern überschätzt ($p < 0,05$).
6. Bei gleicher Echogenität der Lappen ergab sich hinsichtlich der Lokalisation kein Hinweis auf einen Unterschied.
7. Der methodenimmanente Fehler s_{err} (= $\hat{\sigma}_{err}$), der im Falle relativer Differenzen gleichbedeutend ist mit dem Inner-Subjekt-Variationskoeffizienten CV (siehe Punkt 2.4.2.2 in „Material und Methoden“), liegt bei 11% für die Knoten (von den kleinen zu den großen Knoten von 15% auf 8,2% abnehmend) und bei 7,3% für die Lappen.
8. Der reine untersucherbezogene Messfehler („random observer error“ nach Tong et al.) reicht von 0,7% bis 19% für die Knoten und von 1,6% bis 14% für die Lappen.
9. Die mit 95% Wahrscheinlichkeit minimalen, sicher detektierbaren Volumenänderungen betragen für unterschiedliche Untersucher (Interobserver) 55%, 37% und 26% für kleine, mittlere und große Knoten und 23% für die Lappen. Für alle Knoten liegt die Volumenänderung bei 41%. Die Intraobserver-Werte liegen geringfügig darunter.

3.2 Probanden-Studie

3.2.1 3D-Ultraschall-Referenzwerte

Proband	rechts	links	Summe
1	7,45	7,60	15,05
2	14,90	11,10	26,00
3	7,70	6,08	13,78
*4**	*2,30*	*2,95*	*5,25*
*5**	*18,60*	*17,50*	*36,10*
6	11,00	7,40	18,40
7	5,90	5,65	11,55
8	7,00	5,45	12,45
9	7,40	8,80	16,20
*10**	*2,80*	*2,30*	*5,10*
Mittelwert	8,51	7,48	15,99
Streuung	5,08	4,37	9,32
Mittelwert ohne (*)	*8,76*	*7,44*	*16,20*
Streuung ohne (*)	*3,13*	*2,01*	*4,89*

Tab. 8: 3D-Referenzwerte der Schilddrüsenvolumina, differenziert nach Lappen und Gesamtvolumen. Die Volumina, die mindestens eine Streubreite vom Mittelwert abweichen (Proband 4,5 und 10) wurden mit einem Sternchen (*) markiert und die Restgruppe gesondert berechnet (= Normalgruppe). Sie wird in der Reliabilitäts-Analyse gesondert betrachtet.

Fig. 8 illustriert die biometrische Datenverteilung: die Schilddrüsen-Referenzvolumina sind nach Lappen getrennt aufgetragen. Die Gesamtvolumina liegen zwischen 5,1 ml und 36,1 ml. Die rechten Lappen sind im Durchschnitt um ca. 1 ml größer als die linken; ein gepaarter t-Test ergibt p = 0,09, d.h. der Unterschied ist nicht signifikant.

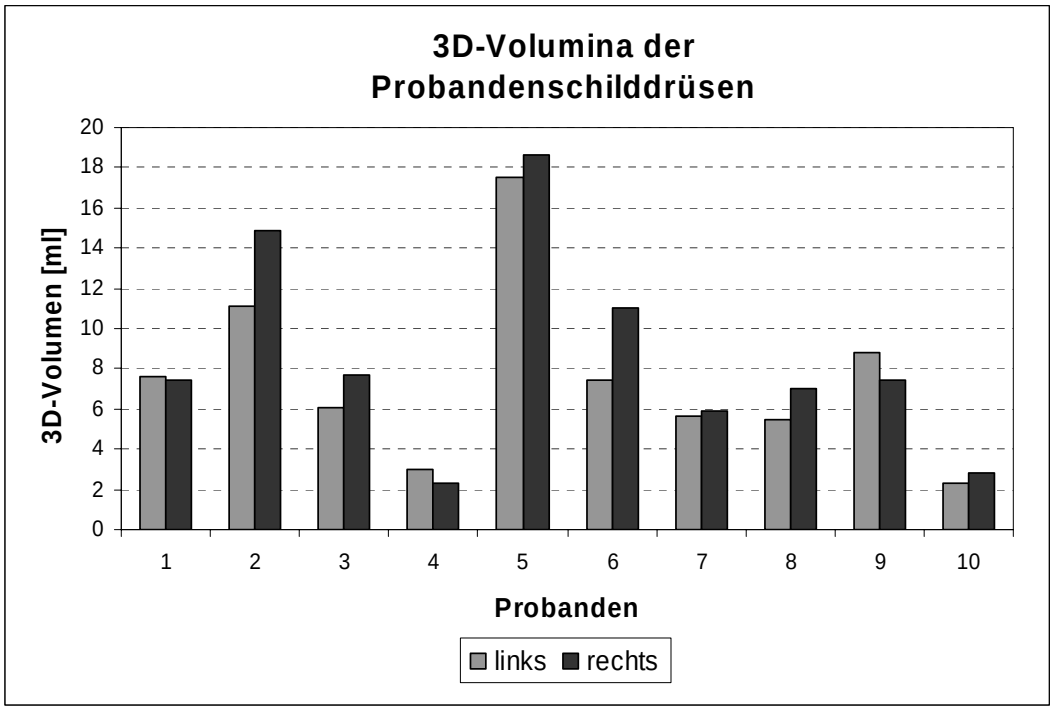
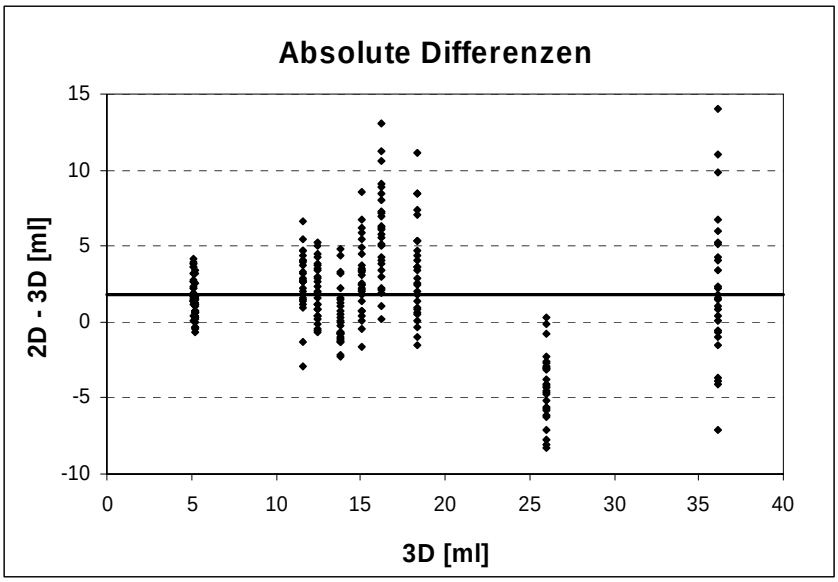


Fig. 8: 3D-Volumina der Probanden-Schilddrüsen, getrennt nach rechtem und linkem Lappen. Die Schilddrüsenvolumina der Probanden 4, 5 und 10 weichen am weitesten vom Mittelwert ab.

Da im Hinblick auf die später durchgeführte Varianzanalyse die Größenunterschiede zwischen den Volumina nicht zu extrem ausfallen sollten, wurden die mit einem Sternchen (*) markierten Schilddrüsenvolumina aus der Hauptgruppe herausgenommen und die Restgruppe als sog. Normalgruppe gesondert berechnet. Die Volumina wurden aufgrund ihrer Daten ausgeschlossen, d.h. sie errechnen sich aus sehr großen (Proband 5) bzw. sehr kleinen (Proband 4,10) Lappen und fallen damit aus dem 1 – σ-Bereich heraus.

3.2.2 Absolute und relative Differenzen



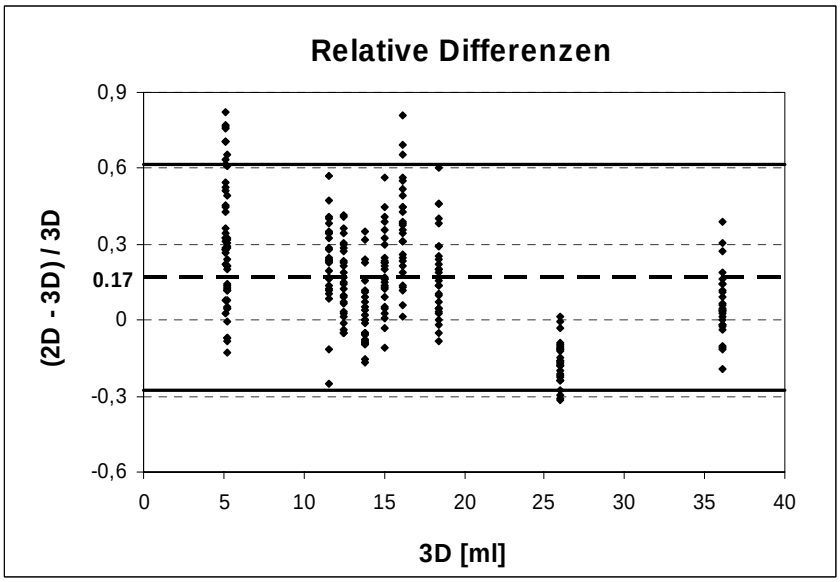
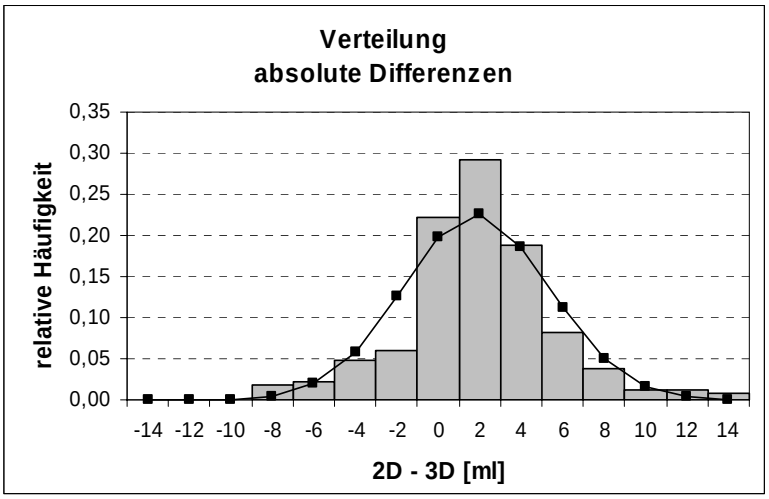


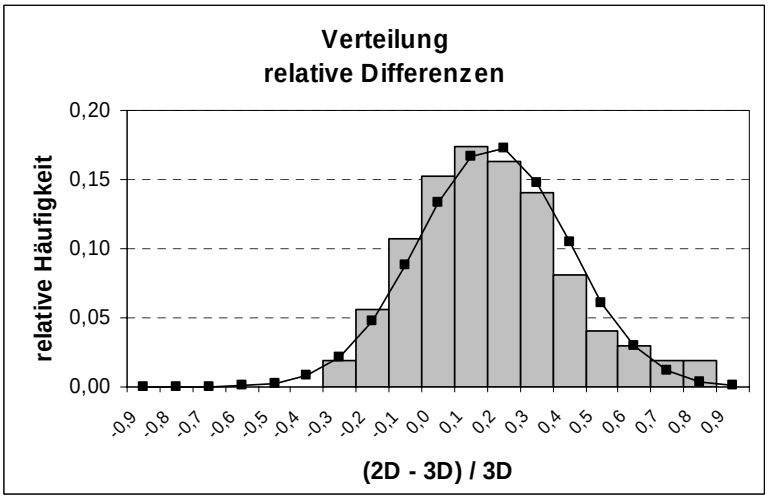
Fig. 9a, b: Absolute und relative Differenzen in Abhängigkeit von den 3D-Referenzmessungen. Die horizontale durchgezogene Linie stellt den Mittelwert aller Messungen dar. Bei den relativen Differenzen ist zusätzlich der +/-95%-Streubereich (Limits of Agreement) eingezeichnet. Die horizontale gestrichelte Linie entspricht dem Mittelwert der Differenzen (+0,17) und zeigt somit den systematischen Fehler an.

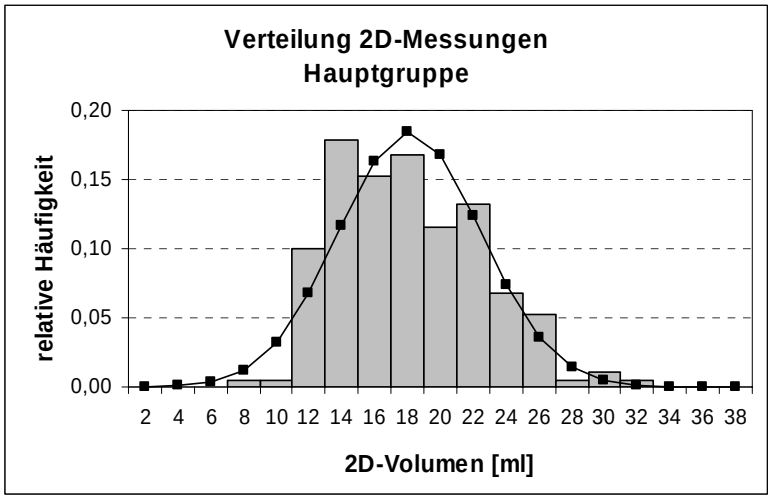
Der Grad der Übereinstimmung der sich aus der untersuchten Methode ergebenden Volumina mit den Referenzvolumina wurde durch das von Bland-Altman [Bland 1986] vorgeschlagene Vorgehen analysiert. Aus Fig. 9a ist, wie bereits bei der Phantom-Studie, die Größenabhängkeit des Messfehlers er-

kennbar: mit ansteigendem Organvolumen nimmt die Streubreite des Fehlers zu. Mit der Transformation zu relativen Differenzen (Fig. 9b) werden die Messfehler über den gesamten Bereich vergleichbar. Nur in diesem Fall ist es sinnvoll, einen horizontalen Streubereich (Limits of Agreement nach Bland-Altman, d.h. das 1,96-fache der Standardabweichung, Normalverteilung vorausgesetzt) anzugeben. Bis auf die Messung an Proband 2 (26 ml) werden alle Volumina deutlich überschätzt, insgesamt um 17%. Dies bezeichnet den systematischen Fehler (Mittelwert der Differenzen) und ist ein Maß für die Richtigkeit der Messmethode.

3.2.3 Häufigkeitsverteilungen der Differenzen







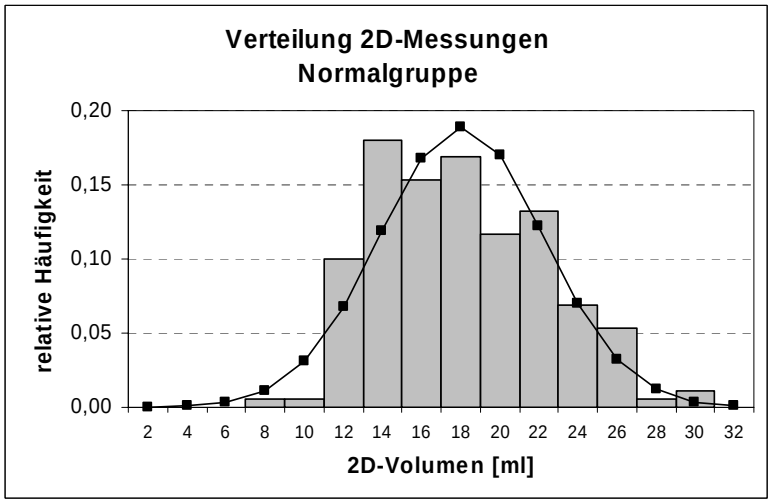


Fig. 10a-d: Verteilung der relativen und absoluten Messungen mit Gauss-Verteilung.

Die Fig. 10 enthält verschiedene Verteilungen: absolute und relative Differenzen aus allen Messwerten und absolute 2D-Messungen für die Gesamt- und Normalgruppe getrennt. Im Hinblick auf die weiter unten zu besprechenden Ergebnisse der Varianzanalyse spricht nichts dagegen, die Daten als normalverteilt anzunehmen.

3.2.4 Übersicht über Messergebnisse und Methodenparameter

MW Streuung	Objekte										
Unter-sucher	1	2	3	4	5	6	7	8	9	10	Intra
1	0,1152 0,0593	-0,1963 0,1004	-0,0010 0,1445	0,1708 0,1368	-0,0030 0,1008	0,0379 0,0947	0,1991 0,0718	0,1604 0,1024	0,2595 0,1561	0,1425 0,1590	0,0885 0,1604
2	0,1371 0,1003	-0,1568 0,0666	-0,0438 0,0589	0,2089 0,0934	0,1667 0,1306	0,1136 0,1679	-0,0949 0,1687	0,0305 0,0884	0,2741 0,0671	0,2928 0,0294	0,0928 0,1732
3	0,1849 0,2052	-0,1095 0,0739	-0,0639 0,0216	0,1606 0,1388	0,0962 0,1207	0,1284 0,1140	0,3671 0,0323	0,2452 0,1634	0,3523 0,0363	0,6516 0,2032	0,2013 0,2386
4	0,1889 0,0833	-0,2324 0,0686	-0,0832 0,0780	0,1683 0,3104	-0,0365 0,1393	0,1663 0,0859	0,2436 0,0777	0,2145 0,1738	0,2755 0,1585	0,4431 0,0908	0,1348 0,2252
5	0,2137 0,2404	-0,1133 0,1073	0,0373 0,0880	0,1441 0,1897	0,0528 0,0670	0,1846 0,0938	0,2395 0,1390	0,1628 0,1397	0,4183 0,1305	0,7039 0,0667	0,2044 0,2447
6	0,1349 0,0777	-0,1722 0,0704	-0,0564 0,0363	0,0844 0,1543	0,0275 0,0152	0,0105 0,0350	0,1752 0,0469	-0,0327 0,0202	0,0870 0,0890	0,2765 0,0000	0,0535 0,1358
7	0,4175 0,1361	-0,1542 0,0557	0,2235 0,1917	0,3346 0,2807	0,1357 0,2455	0,4821 0,1118	0,2173 0,0834	0,2501 0,0843	0,4473 0,2750	0,4817 0,1545	0,2836 0,2416
8	0,1736 0,0589	-0,2306 0,0763	0,1154 0,1498	0,0286 0,1037	-0,0003 0,0377	0,2639 0,1712	0,3437 0,0671	0,3001 0,0639	0,6743 0,1446	0,5248 0,2101	0,2194 0,2733
9	0,1927 0,2590	-0,1263 0,1436	0,1260 0,0951	0,2965 0,3087	0,1082 0,1986	0,2486 0,1812	0,4130 0,1963	0,1390 0,1304	0,4397 0,0618	0,5072 0,2722	0,2345 0,2463
Alle	0,1954 0,1547	-0,1657 0,0862	0,0282 0,1366	0,1774 0,1953	0,0608 0,1316	0,1818 0,1712	0,2337 0,1698	0,1633 0,1419	0,3587 0,1968	0,4471 0,2182	0,1681 0,2289

Tab. 9: Ergebnisse der Probanden-Messungen: Mittelwerte (oben) und Streuungen (unten) der Messungen pro Untersucher und Proband. Relative Differenzen der Gesamtvolumina.

Analog zum Ergebnisteil der Phantom-Studie enthält die Tabelle 9 die Ergebnisse der Messungen pro Untersucher und Proband in Form der Mittelwerte und Streuungen der wiederholten Messungen. Die letzte Spalte enthält die Intraobserver-Variabilitäten, die Werte der letzten Zeile beziehen sich auf die Messungen aller Untersucher pro Objekt. Der untersucherbezogene Intraobserver-Bias reicht von etwa 5% (Untersucher 6) bis 28% (Untersucher 7). Untersucher 6 misst auch am präzisesten (s_{intra} = 14%), Untersucher 8 weist die größte Streuung auf (27%).

Seite	Mittel-wert	s_{inter}	95%-Streubereich (Limits of Agreement)		Std-Fehler	95%-Konf.-Intervall		s_{intra}	s_{err}
rechts	0,2244	0,2812	-0,3267	0,7755	0,0171	0,1908	0,2579	0,2680	0,1700
links	0,1042	0,2415	-0,3692	0,5776	0,0147	0,3566	0,8223	0,2351	0,1549
Summe	0,1681	0,2285	-0,2798	0,6159	0,0139	0,1408	0,1953	0,2200	0,1386

Tab. 10: Mittelwerte und Interobserver-Streuungen mit Streubereichen (Limits of Agreement), Standardfehler und Konfidenzintervall sowie die Intraobserver- und Fehler-Streuungen für beide Lappen und deren Summe. Relative Differenzen.

In Tabelle 10 sind die Ergebnisse der Probanden-Studie zusammengefasst, wobei die Lappen getrennt und das Gesamtvolumen aufgeführt sind. Am 95%-Konfidenzintervall des Mittelwertes ist zu erkennen, dass beide Lappen statistisch signifikant überschätzt werden, das Gesamtvolumen um 17%. Auch die Unterschiede zwischen rechtem und linkem Lappen (22% und 10%) sind signifikant (p < 0,001, t-Test). Die Intraobserver-Streuung s_{intra} beträgt 22% und ist damit nur geringfügig kleiner als der Interobserver-Wert (23%). Die um den Bias bereinigte Mess-Streuung s_{err} ist mit 14% deutlich geringer. Ein F-Test ergibt keinen Hinweis darauf, dass sich die Mess-Streuungen s_{err} der Lappen unterscheiden (F = 1,20, FG = 200,200).

Aus den s-Werten lassen sich wieder die entsprechenden minimalen, sicher detektierbaren Volumenänderungen berechnen (95% Wahrscheinlichkeitsniveau). Für das Gesamtvolumen ergeben sich ΔV_{inter} = 63% und ΔV_{intra} = 61%.

3.2.5 Intraobserver-Variabilität

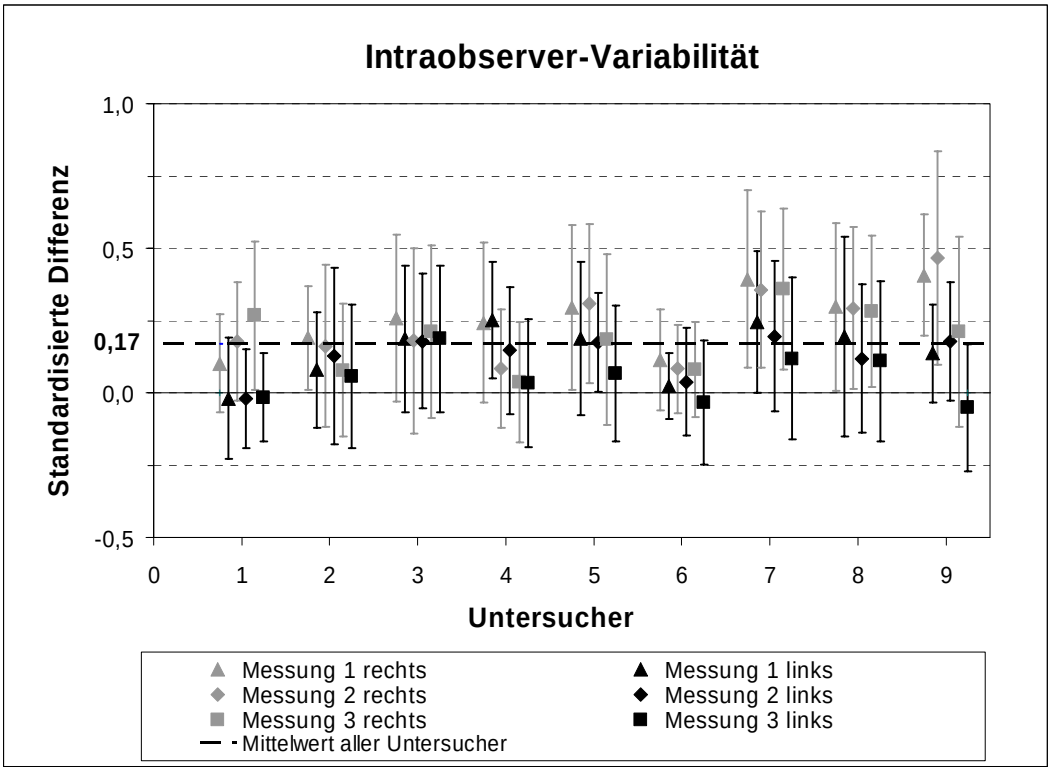


Fig. 11: Intraobserver-Variabilität aller Probanden-Messungen nach Schilddrüsenlappen mit Streubereichen und Gesamtmittelwert (gestrichelte Linie). Der Bias der 2D-Messungen liegt bei +0,17 und zeigt eine systematische Überschätzung der 2D- gegenüber den Referenzvolumina an. Die Fehlerbalken enthalten die Streuung eines Untersuchers über alle Probanden.

In Fig. 11 sind die Mittelwerte und Streuungen aller Messungen pro Untersucher (Intraobserver-Variabilität) differenziert nach Schilddrüsenlappen ablesbar, wobei die Reihenfolgen der Messungen identifizierbar sind. Damit kann beurteilt werden, ob eventuelle Memory-Effekte vorliegen, d.h. die folgenden Messungen der Untersucher am gleichen Probanden von den vorherigen beeinflusst sind. Da sich offenbar kein Muster ablesen lässt, ist ein solcher Effekt nicht erkennbar. Die gestrichelte Linie ist der systematische Fehler (Bias).

Fig. 12 gibt nochmals einen Gesamteindruck über die Intraobserver-Variabilitäten, wobei die Messungen pro Untersucher zusammengefasst sind. Die Mittelwerte aller Untersucher liegen über den Referenzvolumina.

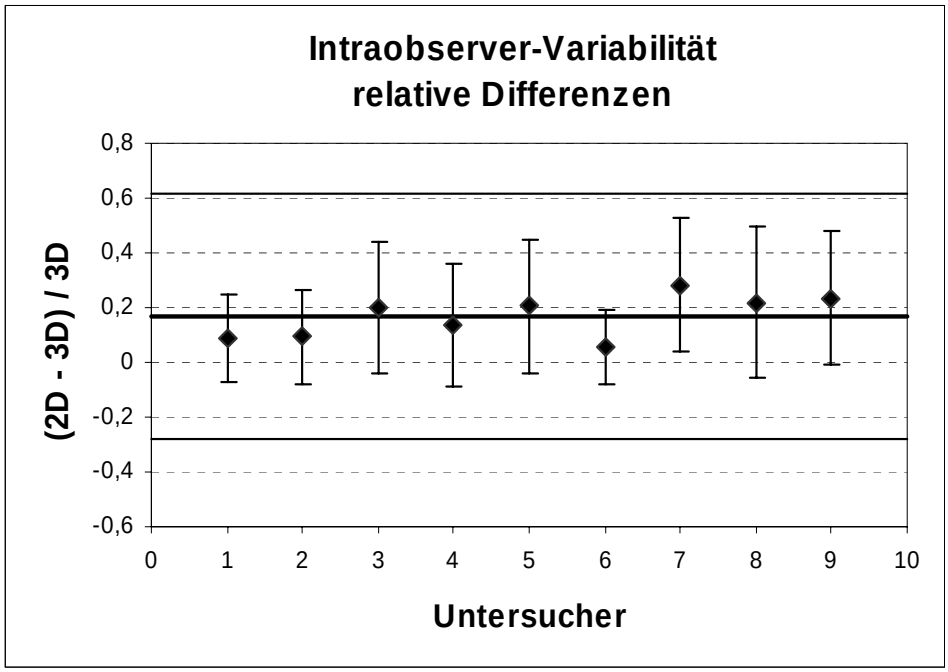


Fig. 12: Intraobserver-Variabilität mit Gesamt-Bias und 95%-Streubereich. Relative Differenzen der Gesamtvolumina, alle Probanden.

3.2.6 Interobserver-Variabilität

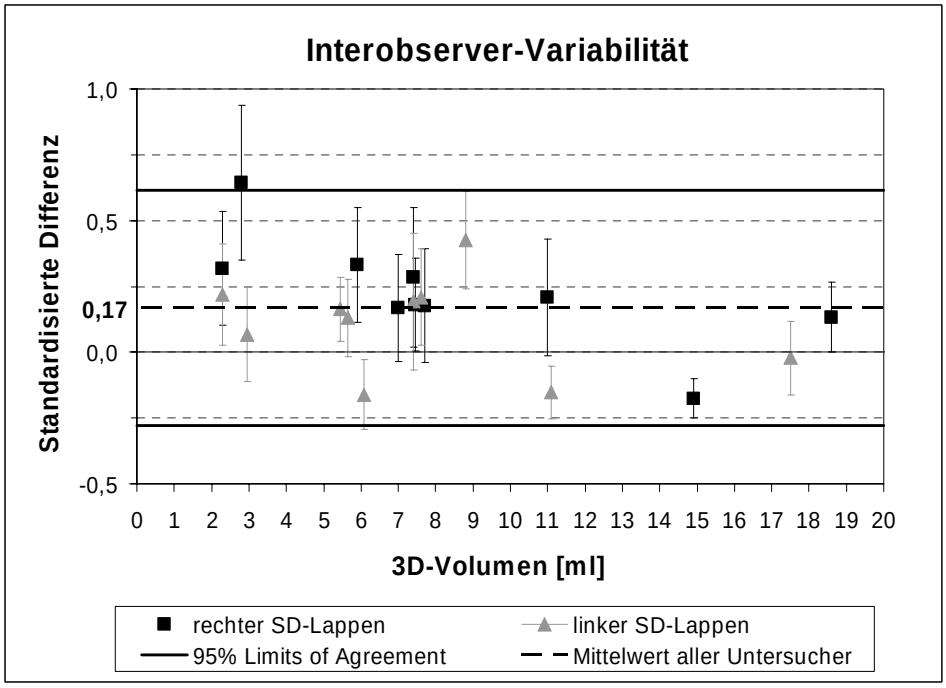


Fig. 13: Interobserver-Variabilität nach 3D-Volumen: standardisierte Differenzen der 2D vs. 3D US-Volumenmessungen von Schilddrüsenlappen bei verschiedenen Untersuchern. Mit Gesamtmittelwert und 95%-Streubereich.

Ähnlich dem Bland-Altman-Plot (Fig. 9b) sind in Fig. 13 alle Daten der Interobserver-Variabilität gegen die Referenzvolumina dargestellt, wobei nach rechtem und linkem Lappen differenziert wird. Zur Orientierung sind wieder Bias und Limits of Agreement eingezeichnet.

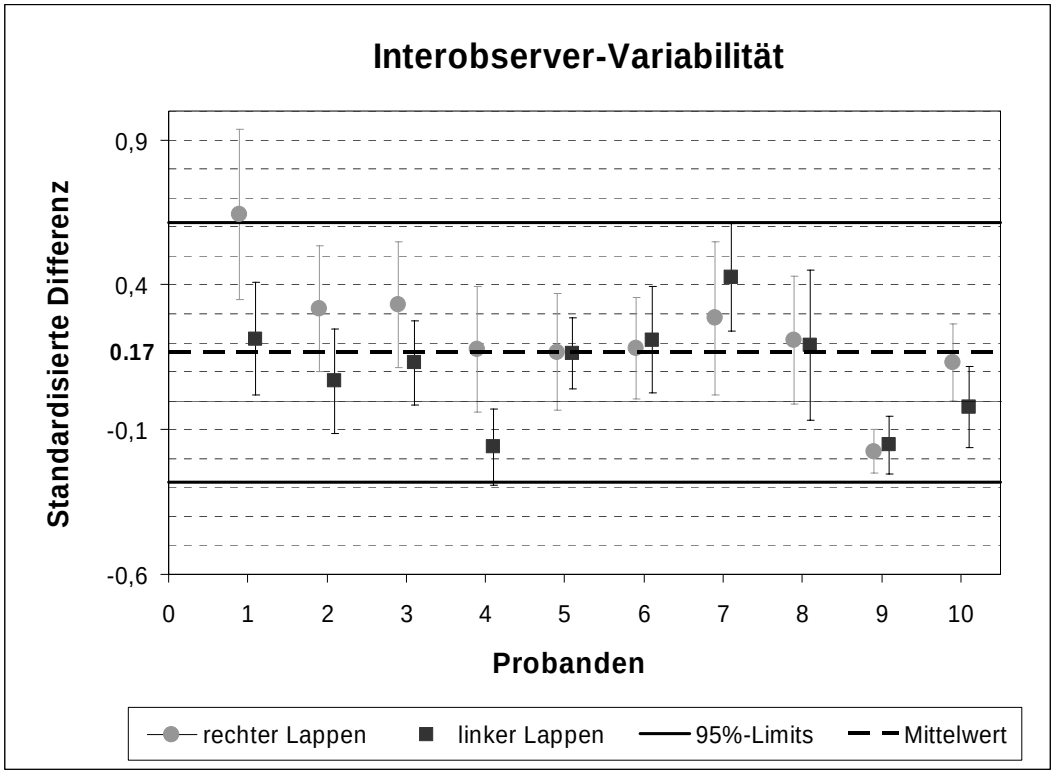


Fig. 14: Interobserver-Variabilität nach Probanden und Schilddrüsengröße: das Gesamtvolumen der Schilddrüsen nimmt von links nach rechts zu.

Fig. 14 illustriert die Interobserver-Variabilität nach Probanden und größensortiert nach dem Schilddrüsenvolumen, jeweils getrennt für beide Schilddrüsenlappen. Mit aufsteigender Probanden-ID nimmt das Gesamtvolumen zu. Die horizontale gestrichelte Linie markiert wieder den Bias der 2D-Messmethode von +0,17, die horizontalen durchgezogenen Linien markieren das 95%-Konfidenzintervall. Erwartungsgemäß liegen nicht alle Datenpunkte innerhalb der Intervallgrenzen. Bei dem vorgegebenen Intervall sind 5% der Messwerte außerhalb der Grenzen zu erwarten.

3.2.7 Zufälliger Untersucher-Fehler (Random Observer Error)

Analog zur Phantom-Studie und weil die Ergebnisse der Probanden-Studie mit den Literaturdaten verglichen werden sollen (vgl. Tong 1998), wurde auch der zufällige Untersucherfehler (Random Observer Error) als die mittlere Streuung der Mehrfachmessungen der Probanden bezogen auf einen Untersucher berechnet (vgl. Punkt 2.4.2.1 in "Statistische Methoden"). Sie sind in Fig. 15 dargestellt. Die Werte bewegen sich zwischen 5,5% und 18%.

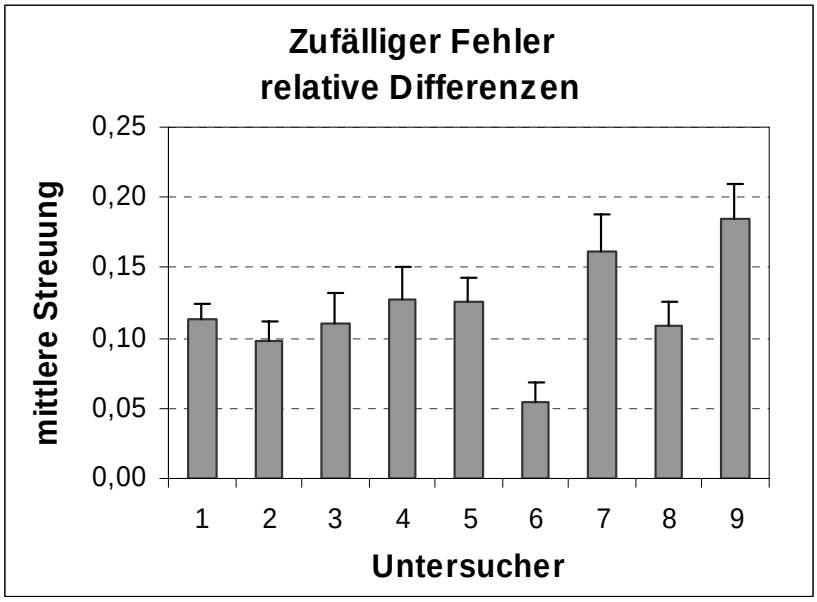


Fig. 15: Zufälliger Fehler (Random Observer Error) der relativen Differenzen mit Standard-Fehlerbalken.

3.2.8 Multivariate Reliabilitätsanalyse

Wie im Methodenteil ausführlich erläutert, wurden die 2D-Ultraschall-Messdaten mit dem von Eliasziw et al. [Eliasziw 1994] vorgeschlagenen Verfahren analysiert. Weil die Ergebnisse stark von der zugrunde liegenden Größenverteilung abhängen, wurde die Analyse der Absolutwerte mit allen Probanden und zusätzlich mit einer Normalgruppe (s.o.) durchgeführt. Außerdem wurde es auf die relativen Volumina zum 3D-Ultraschallverfahren mit allen Probanden angewandt (vgl. Tong 1998).

Fig. 16 zeigt die systematischen Untersucher-Fehler, die einen Eindruck über die unterschiedlichen Variabilitäten der beiden Gruppen geben.

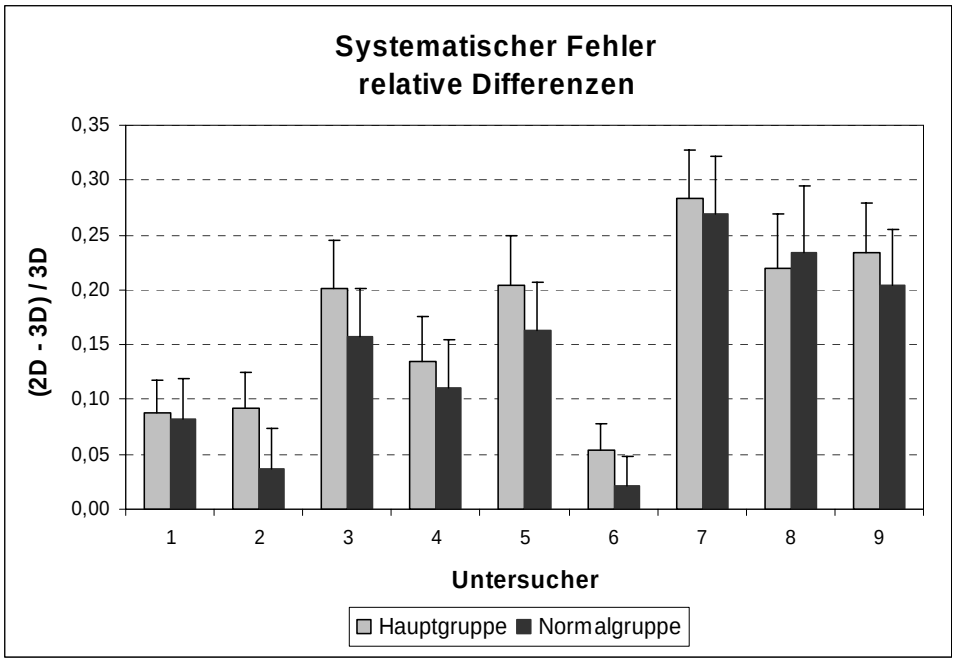


Fig. 16: Systematischer Fehler der relativen Differenzen mit Standard-Fehlerbalken.

Der systematische Fehler entspricht eigentlich dem Intraobserver-Fehler und damit der Intraobserver-Variabilität. In der Literatur [Tong 1998] wird jedoch bisweilen eine andere Form der graphischen Darstellung gewählt (Balkendia-

gramme). Anhand dieser Variante können die Daten besser mit den Literaturergebnissen verglichen werden. Hier wird sie benutzt, um die Unterschiede zwischen der Haupt- und Normalgruppe zu verdeutlichen. Die größere Homogenität der Normalgruppe wird so ersichtlich.

Die Ergebnisse der multivariaten Reliabilitätsanalyse sind in Tab. 11 zusammengefasst, getrennt nach den beiden Lappen und dem Gesamtvolumen.

	2D / 3D			2D (Hauptgruppe)			2D (Normalgruppe)		
	rechts	**links**	**Summe**	**rechts**	**links**	**Summe**	**rechts**	**links**	**Summe**
MS_R	0,3249	0,1650	0,1839	16,4357	7,5629	35,1049	17,9464	7,3877	35,4791
MS_S	1,1082	0,8576	0,7861	676,0758	521,9400	2246,6193	114,3752	170,6916	381,6850
MS_{RS}	0,0496	0,0322	0,0291	2,8795	1,9613	6,5403	2,6376	1,8856	6,0364
MS_{err}	0,0289	0,0244	0,0192	1,9386	1,5953	5,3444	1,8205	1,1791	3,9174
$\hat{\sigma}_R^2$	0,0092	0,0044	0,0052	0,4519	0,1867	0,9522	0,7290	0,2620	1,4020
$\hat{\sigma}_S^2$	0,0392	0,0306	0,0280	24,9332	19,2585	82,9659	4,1384	6,2521	13,9129
$\hat{\sigma}_{RS}^2$	0,0069	0,0026	0,0033	0,3136	0,1220	0,3986	0,2724	0,2355	0,7063
$\hat{\sigma}_{err}^2$	0,0289	0,0244	0,0192	1,9386	1,5953	5,3444	1,8205	1,1791	3,9174
$\hat{\sigma}_{tot}^2$	0,0842	0,0620	0,0557	27,6373	21,1625	89,6611	6,9603	7,9287	19,9387
SEM_{tot}	0,2901	0,2490	0,2360	5,2571	4,6003	9,4690	2,6382	2,8158	4,4653
SEM_{intra}	0,1700	0,1563	0,1386	1,3923	1,2631	2,3118	1,3492	1,0859	1,9792
SEM_{inter}	0,2121	0,1773	0,1663	1,6444	1,3799	2,5875	1,6798	1,2948	2,4547
$\hat{\rho}_{intra}$	0,6567	0,6060	0,6549	0,9299	0,9246	0,9404	0,7384	0,8513	0,8035
$\hat{\rho}_{inter}$	0,4658	0,4930	0,5034	0,9022	0,9100	0,9253	0,5946	0,7885	0,6978
ΔV_{intra}	0,4709	0,4330	0,3840	3,8568	3,4987	6,4037	3,7374	3,0078	5,4825
ΔV_{inter}	0,5874	0,4911	0,4607	4,5550	3,8222	7,1674	4,6531	3,5867	6,7997

Tab. 11: Ergebnisse der multivariaten Reliabilitätsanalyse für die 2D-Ultraschallmethode (alle Probanden und Normalgruppe) und für die relativen Differenzen der beiden Methoden (2D- und 3D-Ultraschall, alle Probanden). Die Werte in den ersten vier Zeilen sind das Ergebnis einer zweifaktoriellen ANOVA (MS = Mean Sum of Squares).

Weil die Objekt-Varianz im Verhältnis zur Observer-Varianz in der Hauptgruppe am größten ist, weisen diese Reliabilitätskoeffizienten $\hat{\rho}_{intra}$ und $\hat{\rho}_{inter}$ die größten Werte auf (0,94 und 0,92). Realistischer in Bezug auf Vergleichbarkeit mit anderen Studien sind sicher die Werte der Normalgruppe (0,80 und 0,70). Die Transformation auf die Referenzvolumina beeinflusst stark den Quotienten der

Objekt- zu Untersucher-Varianz, daher werden hier nur Werte von 0,65 und 0,50 erreicht.

Für die 2D-Volumina der Normalgruppe ergeben sich die Werte ΔV_{intra} = 5,5 ml und ΔV_{inter} = 6,8 ml. Das bedeutet, dass bezogen auf ein mittleres Volumen von ca. 16 ml Volumenänderungen gemessen durch denselben Untersucher erst ab einer Größenordnung von 34% als solche bezeichnet werden können. Bei verschiedenen Untersuchern liegt der Wert bei 43%. Betrachtet man die Hauptgruppe, so wird gerade bei sehr kleinen Schilddrüsenvolumina deutlich, dass eine Aussage zur detektierbaren Volumenänderung nicht möglich ist.

Die folgende Tabelle 12 enthält als Beispiel die Ergebnisse der Varianzanalyse (SPSS) für die Normalgruppe. In der letzten Spalte kann abgelesen werden, ob die Faktoren einen signifikanten Einfluss auf das Ergebnis haben. Die Signifikanz-Wahrscheinlichkeit von 0,27 in der Zeile für die Interaktion Proband – Observer zeigt, dass es keinen Hinweis auf eine überzufällige solche Interaktion gibt, d.h. die Untersucher neigen nicht dazu, etwa kleine Objekte überdurchschnittlich kleiner zu beurteilen.

Tests der Zwischensubjekteffekte						
Abhängige Variable: 2D						
Quelle		Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Signifikanz
Konstanter Term	Hypothese	66100,8640	1	66100,8640	27,6029	0,00316905
	Fehler	12158,0568	5,0771	2394,7064		
Observer	Hypothese	210,2896	8	26,2862	3,3057	0,0054435
	Fehler	318,0711	40	7,9518		
Proband	Hypothese	11881,8599	5	2376,3720	298,8479	1,4938E-30
	Fehler	318,0711	40	7,9518		
Obs * Prob	Hypothese	318,0711	40	7,9518	1,1650	0,26501615
	Fehler	737,1796	108	6,8257		

Tab. 12: Ergebnisse einer SPSS-ANOVA-Beispielberechnung für die Normalgruppe.

3.2.9 Zusammenfassung der Probanden-Ergebnisse

Die Messergebnisse von 9 Untersuchern an 10 Probanden-Schilddrüsen mittels 2D-Ultraschall gegenüber 3D-Ultraschall-Referenzmessungen lassen sich wie folgt zusammenfassen:

1. Die Messungen der Schilddrüsenvolumina in vivo ergeben einen statistisch signifikanten Bias von +17%.
2. Die mittleren Streuungen s_{inter} bzw. s_{intra} liegen bei 23% (22%) für die gesamte Schilddrüse, differenziert nach rechtem und linkem Lappen betragen sie 28% (27%) und 24% (24%). Die Intraobserver-Streuungen liegen damit nur geringfügig unter den Interobserver-Werten.
3. Ein t-Test ergab einen statistisch signifikanten Unterschied der Mittelwerte des rechten (22%) und linken (10%) Schilddrüsenlappens.
4. Die Mess-Streuung s_{err} des Untersuchungsverfahrens, die wieder in etwa dem Inner-Subjekt-Variationskoeffizienten gleichzusetzen ist, liegt bei 14%.
5. Die Werte des reinen Untersucher-Messfehlers („random observer error“ nach Tong et al.) bewegen sich zwischen 5,5% und 18%.
6. Bei den Probanden-Schilddrüsen betragen die mit 95% Wahrscheinlichkeit minimalen, sicher detektierbaren Volumenänderungen 63% bei verschiedenen Untersuchern (Interobserver) und 61% bei nur einem Untersucher (Intraobserver).
7. Die bei der multivariaten Reliabilitätsanalyse errechneten Reliabilitätskoeffizienten $\hat{\rho}_{intra}$ und $\hat{\rho}_{inter}$ weisen mit 0,94 bzw. 0,92 für die Hauptgruppe die größten Werte auf. Die Werte für die Normalgruppe liegen bei 0,80 und 0,70. Nach Transformation auf die Referenzvolumina werden nur Werte von 0,65 und 0,50 erreicht.

3.3 Vergleich der Phantom- und Probanden-Ergebnisse

Betrachtet man die Ergebnisse der Phantom- und der Probanden-Studie nebeneinander, muss berücksichtigt werden, dass ein Vergleich der Studien nur bedingt zulässig ist. Dennoch lassen sich Parameter gegenüberstellen, die Aussagen zu Richtigkeit und Präzision beider Studien in der Zusammenschau ermöglichen. Verglichen werden können allerdings nur Objekte gleicher Größenordnung (Schilddrüsenlappen), wobei beim Phantom aufgrund der beschränkten Objektzahl keine zuverlässigen Schätzungen möglich sind:

Streuung	Phantom Lappen	Probanden Lappen rechts / links	Probanden Gesamtvol.
s_{inter}	8,3%	28% / 24%	23%
s_{intra}	8,7%	27% / 24%	22%
s_{err}	7,3%	17% / 15%	14%

Tab. 13: Interobserver-, Intraobserver- und reine Mess-Streuungen der Phantom- und Probanden-Studie.

Aus den Werten wird ersichtlich, dass die Inter- und Intraobserver-Streuungen bei den Probanden deutlich über den Streuungen der Lappen in der Phantom-Studie liegen. Der Fehler der Messmethode, also der Inner-Subjekt-Variationskoeffizient, fällt bei Verwendung des Phantoms mit 7,3% ins Gewicht, in vivo ist mit einem methodenimmanenten Fehler von 14% zu rechnen.

Beiden Studien ist gemeinsam, dass die Schilddrüsenlappen im 2D-Ultraschall signifikant *über*schätzt werden. Beim Phantom liegt der Bias bei 5,0%, bei den Probanden beträgt er 17%.

Der aus Gründen der Vergleichbarkeit mit der Literatur ermittelte untersucherbezogene Messfehler (Random Observer Error) zeigt für die Lappen des Phantoms eine Spannweite von 1,6% bis 13,7%, bei den Probanden bewegt er sich zwischen 5,5% und 18%.

4. DISKUSSION

Beim Studium der einschlägigen Literatur ist festzustellen, dass die Genauigkeit von Messmethoden nicht einheitlich angegeben wird, was Vergleiche erschwert. Die Palette der Maßzahlen reicht von Spannweiten über mittlere Streuungen bis zu Variationskoeffizienten, wobei letztere noch auf verschiedene Weisen berechnet werden können. Selbst die Begriffe Interobserver-, Intraobserver- und Inner-Subjekt-Variation werden gelegentlich nicht eindeutig voneinander abgegrenzt. Die für den Anwender wohl wichtigste Größe ist die Interobserver-Variabilität, die, wie im Methodenteil definiert, im Zweifelsfall als konservative Schätzung zur Beurteilung herangezogen werden sollte. Dabei ist zu empfehlen, die Berechnungen mit normierten Werten durchzuführen, da dies die Abhängigkeit des Fehlers von der Objektgröße weitgehend eliminiert. Wie im Methodenteil ausgeführt, sind die Ergebnisse dann mit den Variationskoeffizienten vergleichbar. Wird, wie beim Konstrukt des "random observer errors", über Streuungen statt über Varianzen gemittelt, ergeben sich definitionsbedingt kleinere Maßzahlen der Variabilität. Daher wurde diese Größe separat berechnet.

Nicht selten basieren Angaben zum Methodenvergleich noch auf Korrelations- und Regressionsparametern. Auf solche Berechnungen wurde verzichtet, weil sich inzwischen die Erkenntnis durchsetzt, dass sie methodisch nicht adäquat sind.

In einigen wenigen Studien wird ein varianzanalytisches Verfahren verwendet, um Variabilitäten und Reliabilitätskoeffizienten zu schätzen. Es ist bereits darauf hingewiesen worden, dass letztere als absolute Größen betrachtet nur eine beschränkte Aussagekraft haben, weil sie stark von der Verteilung der Objektgrößen abhängen. Aus diesem Grund lassen sich Studien anhand von Reliabilitätskoeffizienten nur vergleichen, wenn die Messungen an ähnlich verteilten Objekten durchgeführt wurden.

Des Weiteren wurde das von Eliasziw eingeführte Konzept der minimal detektierbaren Größenunterschiede weiterentwickelt. Es ermöglicht dem Untersucher, die Wahrscheinlichkeit von falsch positiven Befunden abzuschätzen, wenn die Variabilität der Methode bekannt ist.

4.1 Phantom-Studie

Durch die hier gewählten Methodenparameter wird anhand des Schilddrüsenphantoms, also quasi unter Idealbedingungen, die Frage geklärt, welche Fehlerquellen in das Verfahren eingehen. Am Schilddrüsenphantom sollten Messungen an Objekten unterschiedlicher Größenklassen durchgeführt werden. Quantifiziert werden die Intra- und Interobserver-Variabilität bei Schilddrüsenlappen und bei Schilddrüsenknoten unterschiedlicher Größe und Echogenität, der methodenimmanente Messfehler sowie klinisch relevante Volumenänderungen. Ein Blick in die Literatur zeigt, dass bisher nur wenige Daten zu diesem Thema publiziert wurden, weshalb eine Einordnung der Ergebnisse in einen größeren Kontext nur begrenzt möglich ist.

Intra- und Interobserver-Variabilität

Bei den Messungen der Knoten- als auch der Lappenvolumina konnte ein statistisch signifikanter Bias festgestellt werden. Interessant ist dabei, dass die Lappenvolumina im Durchschnitt um 5,0% von den Untersuchern *über*schätzt werden, während alle Knotenvolumina systematisch *unter*schätzt werden (9,6% für kleine, 8,0% für mittelgroße bis 5,0% für große Knoten). Als Maßzahlen für die Präzision reichen die zugehörigen Inter- bzw. Intraobserver-Streuungen der Knoten s_{inter} und s_{intra} von 20% (18%) bis 9,3% (8,5%). Bezogen auf alle Knoten beträgt der Bias -7,5%, die mittleren Streuungen s_{inter} bzw. s_{intra} liegen bei 15% (14%). Differenziert nach der Knotengröße wird also deutlich, dass mit zunehmender Objektgröße der Bias ansteigt, die Streubreite der relativen Messungen abnimmt. Durch die Transformation zu relativen Differenzen fällt bei kleineren Volumina der Fehler etwas stärker ins Gewicht.

In der Literatur wird eine Studie zur 2D-Schilddrüsenvolumetrie mittels Ultraschall beschrieben, bei der die Richtigkeit der Methode an einem nach einem Autopsiepräparat modellierten Schilddrüsenphantom aus gewebeäquivalentem Material getestet wurde [Szebeni 1992]. Diese Untersuchungen bezogen sich jedoch nur auf das Gesamtvolumen der Schilddrüse. Unter Verwendung des konventionellen Ellipsoidmodells ergaben die Messungen eine systematische Überschätzung des Schilddrüsenvolumens von 17%. Die erhebliche Differenz zu dem Bias der vorliegenden Studie lässt sich dadurch erklären, dass das aktuelle Phantom exakt die geometrische Form eines Ellipsoids darstellt, während bei Szebeni et al. die irreguläre Form einer menschlichen Schilddrüse im Phantom abgebildet wurde. Die daraus resultierenden Messungenauigkeiten und die Inkongruenz von Modell und geometrischer Modellvorstellung führen somit bei mehreren Messungen zu einer Zunahme des systematischen Fehlers.

Um die Variationsvielfalt humaner Schilddrüsen zu kompensieren, führten Brunn et al. [Brunn 1981] einen empirisch ermittelten Korrekturfaktor (f = 0,479) in die Ellipsoidformel ein. Dadurch ließe sich der systematische Fehler im Verhältnis zu dem hier benutzten Faktor f = 0,5 zwar etwas reduzieren, aber nicht vollständig eliminieren. Untersucht man Patienten oder Probanden im Rahmen von Studien, so ist mit einem ähnlichen Bias wie bei Szebeni et al. zu rechnen, da in der Realität die Schilddrüsenlappen ebenfalls vom ellipsoiden Modell abweichen. Dieser Aspekt wird im Diskussionspunkt 4.2 (Probanden-Studie) noch ausführlicher erörtert.

Bei den untersucherbezogenen Messungen der Knotenvolumina (Fig. 3) fällt eine zum Teil recht hohe individuelle Schwankungsbreite der Messergebnisse auf (Intraobserver-Variabilität). Bei Untersucher 1 und 8 differieren die Mittelwerte der Differenzen stark zwischen rechtem und linkem Lappen, lassen also eine deutliche Diskrepanz bei der Richtigkeit erkennen, während andere Untersucher (z.B. 2 und 4) sehr konsistente Ergebnisse liefern. Unpräzise aufgrund ihrer großen Streubreite sind vor allem die Messungen von Untersucher 7 und 9. Dagegen vereint sich bei Untersucher 4 ein geringfügiger systematischer

Fehler mit einer geringen Standardabweichung, was auf eine adäquate und sehr zuverlässige Reproduktion des tatsächlichen Volumens hindeutet. Insgesamt weisen die Ergebnisse der Intraobserver-Fehler ein sehr inhomogenes Verteilungsmuster auf. Dabei zeigt sich, dass die Mess(un)genauigkeit einzelner Untersucher nicht unbedingt mit deren klinischer Erfahrung korreliert. Dieser Punkt wird im Diskussionsteil „Probanden-Studie“ noch einmal aufgegriffen. Von wenigen Ausnahmen abgesehen ist die Konsistenz der Lappenmessungen erwartungsgemäß deutlich höher als bei den Knotenmessungen. Auch liegt hier der systematische Fehler der einzelnen Untersucher etwa in der Größenordnung der Streuungen.

Vergleicht man die systematischen Abweichungen der Knotenmessungen unterschiedlicher Echogenität (-5 dB und -10 dB) miteinander, lassen sich auch hier quantitative Aussagen machen. Die statistische Auswertung der Ergebnisse lässt eine, wenn auch nur knapp über dem Signifikanzniveau liegende Tendenz erkennen, die echoreicheren Knoten etwas genauer zu messen. Warum gerade die kontastärmeren Läsionen einen größeren systematischen Fehler aufweisen, bleibt unklar. Möglicherweise neigen Untersucher bei klar abgegrenzten Objekten dazu, die Begrenzungspunkte der Durchmesser etwas enger zu wählen als bei weniger klar definierten Grenzen. Der erwartete Einfluss der Echogenität ließ sich jedenfalls nicht bestätigen. Auch bei den Untersuchungen von Brauer et al. [Brauer 2005] war durch den Faktor Echogenität keine Beeinflussung des Variationskoeffizienten zu erkennen. Bei der Analyse der Messvarianzen der -5 dB- und -10 dB-Knoten war jedoch im F-Test eine knapp an der Testschranke liegende größere Streubreite bei den Volumendaten der kontrastreicheren Läsionen festzustellen, was zumindest auf einen bestehenden Unterschied hindeutet. Zu erklären ist dieser Befund am ehesten dadurch, dass sich gerade die echoreicheren Läsionen aufgrund der niedrigeren Impedanzschwelle schwerer vom umgebenden Parenchym abgrenzen lassen.

Aus Gründen der Vergleichbarkeit mit der Literatur wurde nach Tong et al. [Tong 1998] der zufällige, vom Bias unabhängige Untersucherfehler (Random

Observer Error), also die mittlere reine Mess-Streuung pro Untersucher, berechnet. Hierbei ergaben sich für die Knoten Werte von 0,7% bis 19,5%, für die Lappen reichten die Werte von 1,6% bis 13,7%. Die korrespondierenden Werte von Tong et al. liegen etwas höher, nämlich zwischen ca. 7% und ca. 24%. Dieser Vergleich ist nur wegen des ähnlichen Studiendesigns statthaft. Einschränkend muss jedoch angemerkt werden, dass bei Tong et al. Prostata-Volumina unter In-vivo-Bedingungen untersucht wurden. Somit sind die höheren Messwerte im Gegensatz zur vorliegenden „In-vitro-Situation“ erwartungskonform.
Aus diesen Berechnungen lässt sich weiterhin ableiten, dass bei den isoechogenen Lappen zwar individuelle Variationen erkennbar sind, aber keine Präferenz bezüglich der Lokalisation, d.h. die Untersucher zeigen keine Tendenz, eine Seite genauer zu messen.
Der Intraobserver-Bias erstreckt sich bei den Knoten von -22,3% bis +8,7% (Median -5,3%), für die Lappen reichen sie von -8,1% bis +18,4% (Median 5,3%). Bei Tong et al. (dort als "systematic observer error" aufgeführt) wird ein Schwankungsbereich von ca. -2% bis ca. +38% angegeben, was wieder der In-vivo-Situation für Prostata-Volumina zuzuschreiben ist.

In der Vergangenheit sind bereits mehrfach in klinischen Studien die Volumenänderungen von Schilddrüsenknoten ohne Pharmakotherapie bzw. als Therapie-Response auf Levothyroxin- oder Iodid-Gabe mittels 2D-Ultraschall untersucht worden [Hansen 1979, Cheung 1989, Papini 1993, Kuma 1994, Mainini 1995, Papini 1998, Quadbeck 2002]. Bisher liegen dazu jedoch keine randomisierten, plazebokontrollierten Studien mit ausreichender Fallzahl vor. Aus diesem Grund wurde 2004 eine multizentrische, randomisierte, doppelblinde, vierarmige Studie, die sog. LISA-Studie, in Deutschland aufgelegt [Grussendorf 2005].
Klinisch relevanter sind allerdings Volumenänderungen unter Pharmakotherapie bei der euthyreoten diffusen Struma, insbesondere der diffusen Iodmangelstruma des jungen Erwachsenenalters. In einer ganzen Reihe von Studien konnte nämlich nachgewiesen werden, dass bei dieser Form der Schilddrüsenerkrankung durch Iodid, Levothyroxin oder Kombinationen aus beiden Präparaten eine

Volumenreduktion zu erzielen ist [Hintze 1985, Olbricht 1985, Pfannenstiel 1988, Hintze 1989, Einenkel 1992, Wilders-Truschnig 1993, Grussendorf 1996, Rönnefarth 1996, Förster 1998, Klemenz 1998, Kreißl 2001]. Da Fehlerabschätzungen allerdings realistischer bei Probanden in vivo vorgenommen werden können, wird auf diesen Punkt bei der Diskussion der Probanden-Ergebnisse näher eingegangen.
Auch die Volumenreduktion nach einer Radioiodtherapie bei Patienten mit fokaler Schilddrüsenautonomie war bereits Gegenstand mehrerer Untersuchungen [Luster 1995, Dederichs 1996, Le Moli 1999, Nygaard 1999, Reinhardt 2002b]. Allerdings ist die klinische Interpretation einer Volumenzu- oder -abnahme von Knoten im 2D-Ultraschall problematisch und fehleranfällig. Bei den meisten der zitierten Studien wurde das Knotenvolumen durch mehrere Untersucher bestimmt. Somit setzen sich die Limitationen des Verfahrens aus dem Fehler der Messmethode und der Subjektivität bzw. der Erfahrung des Untersuchers zusammen. Das Hauptproblem für einen Ultraschallanwender besteht darin, 2D-Bilder aus ziemlich willkürlich gewählten Schnittebenen durch ein Organ zu beurteilen und die Informationen geistig in eine dreidimensionale anatomische Struktur zu integrieren. Dieses klinische Procedere kann zu einer erheblichen Interobserver-Variabilität und sogar zu falschen Diagnosen führen [Hamper 1999, Fenster 2001], da Organe durch einfache geometrische Figuren häufig nicht korrekt beschrieben werden.

In der klinischen Praxis ist es wichtig zu wissen, ab welchem Betrag ein Unterschied zweier Messungen an einem Objekt (z.B. Schilddrüsenknoten) auf einer tatsächlichen Volumenänderung beruht oder durch einen Messfehler bedingt und damit rein zufällig sein könnte. Will man beispielsweise den Effekt einer medikamentösen Behandlung auf das Volumen beurteilen, braucht man vor, während und nach der Therapie quantitativ genaue Messwerte, um einen potenziellen Verkleinerungseffekt objektivieren zu können.
Bei der vorliegenden Studie ergibt sich für alle Knoten bei einer Interobserver-Streuung von 15% ein Cut-off von 41% als die mit 95% Wahrscheinlichkeit minimale, sicher detektierbare Volumenänderung. Differenziert nach der Knoten-

größe betragen die Werte 55% für die kleinen, 37% für die mittleren und 26% für die großen Knoten. Die jeweiligen Intraobserver-Werte liegen etwas darunter. Bei den meisten klinischen Studien, die den Therapieeffekt von Schilddrüsenhormonen auf intrathyreoidale Herdbefunde untersuchen, wird ein Cut-off von 50% oder mehr zugrunde gelegt [Cheung 1989, Reverter 1992, La Rosa 1995, Papini 1998, Zelmanovitz 1998, Wemeau 2002]. Auf geringere Volumenabnahmen (≤ 50%) wurde in den meisten Studien nicht eingegangen [Gharib 1987, Reverter 1992, Mainini 1995, Papini 1998, Zelmanovitz 1998]. In einer kürzlich von Brauer et al. [Brauer 2005] veröffentlichten Studie zur Volumenbestimmung von Schilddrüsenknoten nach der Ellipsoidmethode im Ultraschall wird für die Interobserver-Streuung ein Wert von 49% ermittelt. Auch hier lautet die Empfehlung, dass – vergleichbare Streuungen vorausgesetzt – erst Zu- oder Abnahmen von 50% oder mehr als Knotenschrumpfung bzw. -wachstum oder als Therapieeffekt interpretiert werden sollten.
Kritisch sind jene Empfehlungen zu bewerten, die den Cut-off für eine klinisch relevante Volumenänderung bereits bei Werten in der Größenordnung der Streuung oder darunter ansetzen [Hansen 1979, Celani 1990, Gullu 1999, Quadbeck 2002]. In dem Paper von Papini et al. [Papini 1998] wird empfohlen, zur Detektion von Volumenänderungen den Variationskoeffizienten CV zu verwenden. Dies bedeutet, dass eine Irrtumswahrscheinlichkeit von etwa 48% in Kauf genommen wird. Somit ist kritisch anzumerken, dass der von Papini et al. festgelegte Variationskoeffizient von 11,7% ein hohes Fehlerrisiko impliziert.

Den Daten in der Literatur zufolge ist bei Patienten mit unifokaler Autonomie nach einer Radioiodtherapie mit einer Volumenreduktion des Knotens im Ultraschall von etwa 40% zu rechnen [Reiners 1993, Reiners 2002]. Aber auch andere Studien haben in der Folge gezeigt, dass bei jeder Art von funktioneller Schilddrüsenautonomie nach Radioiodtherapie eine Volumenabnahme von etwa 40% erreicht werden kann [Guhlmann 1992, Luster 1995, Dederichs 1996, Le Moli 1999, Nygaard 1999, Reinhardt 2002b].
Die eigenen Daten stützen also die Ergebnisse in der Literatur. Wenn man im Phantom bei den mittleren und großen Knoten (Volumen 0,5 ml und 1,4 ml), die

die häufigsten Herdbefunde bei Patienten darstellen, minimal detektierbare Volumenänderungen von 37% bzw. 26% zugrunde legt, muss man von der Vorstellung ausgehen, dass das Modell eine Idealsituation darstellt, die die Wirklichkeit nicht ganz exakt abbildet. Daher ist bei Volumenbestimmungen in vivo ein etwas größerer Fehler anzunehmen. Ein Interobserver-Cut-off zwischen 40% und 50% kann somit bei der klinische Evaluierung von Volumenänderungen von Schilddrüsenknoten als angemessen gelten. Auf der anderen Seite ist es nahe liegend anzunehmen, dass die Reproduzierbarkeit zwischen zwei Untersuchungen besser wird, wenn die Ergebnisse der vorherigen Untersuchung bekannt sind. Daher sollten dem nächsten Untersucher alle bekannten Daten zu einem Patienten zur Verfügung stehen.

Derzeit läuft in Deutschland die bereits oben erwähnte Multicenter-Studie (LISA-Studie) [Grussendorf 2005] mit dem Ziel, die pharmakologische Behandlung von Schilddrüsenknoten zu optimieren. Um die Qualitätskontrolle während des Studienablaufs sicherzustellen, soll das Phantom zu den teilnehmenden Studienzentren ($n > 50$) gebracht werden, um überall einen hohen Standard für Bildqualität und Volumenmessungen zu gewährleisten.

Das Schilddrüsenphantom kann auch für die periodisch erforderlichen Kontrollen von Ultraschallgeräten verwendet werden. Andere, kommerziell erhältliche Ultraschallphantome, die aus Nylonfäden und „zystischen“, scheibenartigen Objekten bestehen und zur Qualitätskontrolle von Ultraschallgeräten dienen, simulieren Läsionen, die – anders als bei humanen Schilddrüsen – im Vergleich zum umgebenden Parenchym keine Kontrastdifferenz aufweisen. Daher eignen sie sich nicht für Langzeitmessungen im Rahmen der Qualitätskontrolle von Schilddrüsenuntersuchungen. Ein Vergleich dieser Mehrfachmessungen mit einem Referenzbild lässt im Laufe der Zeit keine Kontrastverschlechterung bei dem hier verwendeten Ultraschallscanner erkennen. Der Einsatz des Phantoms bei Langzeit-Qualitätskontrollen gewährleistet qualitativ hochwertige Patientenuntersuchungen, was nach den heutigen Standards noch nicht bindend vorgeschrieben ist. Aktuell finden sich in der Literatur keine Publikationen zu

Fragen der Bildqualität. Zudem wurden bisher keine Studien zur Langzeitstabilität von Ultraschallgeräten veröffentlicht.

Als klinische Konsequenz ergibt sich daraus, das Schilddrüsenphantom als Simulationsobjekt für reale Untersuchungsbedingungen einzusetzen mit dem Ziel, die manuellen Fähigkeiten des einzelnen Ultraschallanwenders zu trainieren. Damit sollen in der klinischen Routine konsistentere Ergebnisse erreicht werden. Dies soll besonders für Untersucher gelten, die noch in der Ausbildung sind. Aber auch für erfahrene Ultraschallanwender ergibt sich die Möglichkeit, Richtigkeit und Präzision ihrer Messungen durch Vergleich mit wahren Volumina zu verifizieren.

Aktuell ist ein wachsendes Interesse an der 3D-Ultraschallvolumetrie festzustellen. Mit dem Phantom lässt sich auf sehr einfache Art die Genauigkeit von 3D-Ultraschallsystemen überprüfen. Sind erst einmal Richtigkeit und Reproduzierbarkeit eines 3D-Systems etabliert, können damit in vivo Referenzwerte bestimmt werden, wenn das wahre Schilddrüsen- oder Herdvolumen unbekannt ist. Dieses Verfahren wurde bei der Probanden-Studie angewandt.

Für das hier vorgestellte neuartige Schilddrüsenphantom ergeben sich somit vielfältige Einsatzmöglichkeiten:

- Mit seiner Hilfe können die individuellen Fähigkeiten von Ultraschallanwendern überprüft werden.
- Durch Simulation der realen Schilddrüsenmorphologie sollen die manuellen Fertigkeiten von Ärzten auf dem Gebiet der Ultraschalldiagnostik in Form von regelmäßigen Übungseinheiten trainiert werden, letztendlich um bessere Ergebnisse in der klinischen Routine zu erreichen und die Qualität von Ultraschallergebnissen insgesamt zu verbessern.
- Es kann zur Qualitätskontrolle bei klinischen Studien eingesetzt werden.
- Als standardisiertes Referenzobjekt für Verlaufsmessungen kann es zur Langzeit-Qualitätskontrolle von konventionellen Ultraschallgeräten dienen.

- Durch den Einsatz von 3D-Ultraschallsystemen können Richtigkeit und Reproduzierbarkeit der 2D-Schilddrüsenvolumetrie bestimmt werden.

An der Klinik für Nuklearmedizin der Universität Würzburg konnte somit ein Phantom kreiert werden, das einfach zu handhaben und jederzeit verfügbar ist, das für den Untersucher eine diagnostische Herausforderung darstellt und den Schwerpunkt auf reproduzierbare Volumenmessungen legt. Dies soll insbesondere für Untersucher gelten, die noch in der Ausbildung sind. Aber auch für in der Sonographie erfahrene Ärzte ergibt sich die Möglichkeit, die Richtigkeit ihrer Messungen durch Vergleich mit den wahren Volumina zu verifizieren. Mit Hilfe des Phantoms konnten unter simulierten Untersuchungsbedingungen zum ersten Mal die Intra- und Interobserver-Variabilität bei der Volumenbestimmung vorgegebener Ellipsoide unterschiedlicher Echogenität sowie der methodenimmanente Messfehler und detektierbare Volumenänderungen arithmetisch erfasst und quantifiziert werden. In Zukunft sollen noch andere Modalitäten einfließen wie z.B. komplexere pathologische Befunde. Um die Feinnadelbiopsie zu simulieren, müsste allerdings ein resistenteres Material gewählt werden, denn die physische Integrität des akutellen Phantoms würde durch diese Manipulation irreversibel verändert werden.

4.2 Probanden-Studie

Die vorliegenden Daten stellen die erste prospektive verblindete klinische Studie dar, die die Evaluierung der Intra- und Interobserver-Variabilität bei der Messung des Schilddrüsenvolumens gesunder Erwachsener mittels 2D-Ultraschall gegenüber 3D-Referenzvolumina zum Ziel hat. Als weitere Einflussfaktoren bzw. mögliche Fehlerquellen werden der methodenimmanente Messfehler und die minimal detektierbaren Volumenänderungen quantifiziert. Die Bestimmung des zufälligen Fehlers nach Tong et al. [Tong 1998] soll wieder dem Vergleich mit Literaturdaten dienen. Aus dem gleichen Grund wurde zusätzliche eine multivariate Reliabilitätsanalyse durchgeführt.

Mit dem Studiendesign wurden bewusst die Untersuchungsbedingungen der täglichen Routinetätigkeit simuliert. Dadurch haben die Ergebnisse einen klaren klinischen Bezug und leisten einen wichtigen Beitrag bei der Anwendung des 2D-Ultraschalls in der Praxis.

Intra- und Interobserver-Variabilität

Betrachtet man die Richtigkeit (accuracy) der hier angewandten 2D-Methode, so zeigen die Bland-Altman-Plots in Fig. 11 ff. eine signifikante Überschätzung der Schilddrüsenvolumina im 2D-Ultraschall von +17% ($p < 0,01$) gegenüber den 3D-Referenzvolumina. Darüber hinaus lässt sich erstaunlicherweise ein statistisch signifikanter Unterschied zwischen den Mittelwerten des rechten (22%) und linken (10%) Schilddrüsenlappens feststellen, d.h. die linken Lappen werden systematisch genauer gemessen. Nachdem sich zwischen den Lappenreferenzvolumina beider Seiten keine statistisch bedeutsame Differenz zeigt, kann über dieses Phänomen nur spekuliert werden. Möglicherweise spielt die Angulierung des Schallkopfes eine Rolle, die vermutlich zu einer besseren Abgrenzung des linken Lappens führt.

Der mittlere Bias von +17% entspicht den Befunden von Brunn et al. [Brunn 1981], die die 2D-Ultraschallmethode an 25 Leichenschilddrüsen evaluierten und eine Abweichung von durchschnittlich 16% feststellten, und mit den Resultaten von Szebeni und Beleznay [Szebeni 1992], die In-vivo-Studien an 40 Probanden durchführten und bei Anwendung des konventionellen Ellipsoidmodells einen systematischen Fehler von +19% ermittelten. Igl et al. [Igl 1981] bezogen bei ihrer Studie an 101 Patienten in die Ellipsoidformel die szintigraphisch bestimmte Längen- und Breitenausdehnung und die sonographisch gemessene Schilddrüsendicke ein und errechneten einen durchschnittlichen Fehler von 14%. Auch eine kürzlich publizierte Studie [Rago 2006] berichtet in einer Subgruppenanalyse (gleichzeitiges Vorliegen intrathyreoidaler Knoten) von einer systematischen Überschätzung des Schilddrüsenvolumens um +10% im 2D- verglichen mit dem 3D-Ultraschall, wobei der prozentuale Messfehler bei den kleineren Volumina höher lag. Dass in der Studie von Lyshchik et al. [Lyshchik 2004a] der systematische Fehler der 2D-Ultraschallvolumetrie im Vergleich zu

den OP-Präparaten nach totaler Thyreoidektomie mit +3,2% so gering ausfiel, ist wohl am ehesten durch den Selektions-Bias zu erklären: die Untersuchung war beschränkt auf 47 Jugendliche mit einem mittleren Alter von 14 Jahren, bei denen wegen Schilddrüsenknoten die Indikation zur Operation gestellt worden war.

Um diesen Bias zu reduzieren, führten Brunn et al. [Brunn 1981] einen empirisch ermittelten Korrekturfaktor $f = 0{,}479$ ein, der in Deutschland weit verbreitet ist und anstelle von $f = \frac{\pi}{6}$ (~0,524) als Bestandteil der standardmäßigen Ellipsoidformel verwendet wird. Basierend auf dieser Publikation benutzt sogar die World Health Organization (WHO) seit einiger Zeit diesen modifizierten Korrekturfaktor zur Berechnung des Schilddrüsenvolumens [Shabana 2006]. Aus Gründen der Vereinfachung wurde hier – analog zur täglichen klinischen Routine – der Korrekturfaktor $f = 0{,}5$ verwendet. Ersetzt man diesen durch den Faktor $f = 0{,}479$ von Brunn, ließe sich der systematische Fehler bei dieser Studie um weitere 4,2% verringern.

Andere Autoren [Knudsen 1999, Nygaard 2002, Reinartz 2002, Van Isselt 2003] wiederum behaupten, dass durch den 2D-Ultraschall das wahre Schilddrüsenvolumen *unters*chätzt werde. Die Angaben reichen von einer quantitativ nicht erfassten geringen Tendenz zur Unterschätzung [Knudsen 1999] bis hin zu einem Bias von -17% im Ultraschall verglichen mit CT-Volumina [Nygaard 2002]. Auch Miccoli et al. [Miccoli 2006] berichten in einer präoperativ an 101 Schilddrüsen mit einem maximalen Volumen von 50 ml durchgeführten Studie zur 2D-Ultraschall-Volumetrie von einer statistisch signifikanten Volumenunterschätzung (88% der Fälle) im Vergleich zu den durch Submersion ermittelten Referenzvolumina nach totaler Thyreoidektomie.

Als Konsequenz daraus wird abgeleitet, dass der ursprüngliche Faktor der Ellipsoidformel $f = \frac{\pi}{6}$ [Knudsen 1999, Nygaard 2002] oder sogar ein noch höherer Faktor $f = 0{,}6$ [Reinartz 2002] richtigere Messergebnisse liefern würden. In die gleiche Richtung zielt eine kürzlich erschienene Studie von Shabana et al. [Shabana 2006], derzufolge sich Schilddrüsenvolumenmessungen bei Verwendung des Korrekturfaktors $f = 0{,}479$ in der Ellipsoidformel statistisch signifikant

von softwarebasierten automatisierten Volumenbestimmungen mittels Computertomographie unterscheiden. Als akzeptable Korrekturfaktoren definieren die Autoren ein Intervall von 0,494 bis 0,554 und empfehlen, für Volumenberechnungen mit der Ellipsoidformel den Mittelwert 0,529 zu verwenden. Ähnlich äußerten sich schon vor Jahren Igl et al. [Igl 1981], die einen empirisch gefunden Mittelwert von f = 0,53 definierten. Die Tatsache, dass dieser so nah am theoretischen Wert $\frac{\pi}{6}$ = 0,524 liege, sei ihrer Meinung nach als Hinweis auf die prinzipielle Gültigkeit des Ellipsoidmodells zu deuten.

Die Daten der vorliegenden Studie legen den Schluss nahe, dass mit einem Korrekturfaktor f = 0,45 die beste Volumennäherung möglich ist. Dabei ist grundsätzlich zu betonen, dass sich durch die Einführung von Korrekturfaktoren nur der systematische Fehler reduzieren lässt, die Streubreiten bleiben unverändert. Weiterhin ist zu bedenken, dass im Falle einer Volumen*unter*schätzung der Fehler dadurch noch vergrößert wird. Einschränkend ist weiterhin zu bemerken, dass die vorliegende Untersuchung an normal großen und gering vergrößerten Schilddrüsen durchgeführt wurde. Bei großen und multinodös veränderten Strumen ist möglicherweise aufgrund einer systematischen Volumen*unter*schätzung ein höherer Korrekturfaktor anzusetzen, was ohnehin von verschiedenen Autoren gefordert wird [Nygaard 2002, Reinartz 2002, Shabana 2006].

Neben dem systematischen Fehler, der auf den Rechenoperationen mit dem konventionellen Ellipsoidmodell basiert, haben die Messungenauigkeiten der Observer einen nicht unerheblichen Anteil an den zum Teil sehr inkonsistenten Ergebnissen in der Literatur. Ob sich jedoch exaktere Volumenbestimmungen – wie von einigen Autoren vorgeschlagen [Igl 1981, Szebeni 1992, Miccoli 2006] – durch „bessere“ mathematische Modelle erreichen lassen, die auch die irreguläre Form von abnormen Schilddrüsen berücksichtigen sollen, erscheint eher fraglich.

Als Inter- bzw. Intraobserver-Streuungen ergeben sich Werte von 23% bzw. 22% für die gesamte Schilddrüse; nach Lappen differenziert betragen sie 28% (27%) für den rechten und 24% (24%) für den linken Schilddrüsenlappen, woraus sich kein signifikanter Unterschied ableiten lässt.
In vivo wurde für die um den Bias bereinigte Mess-Streuung s_{err} des 2D-Ultraschallverfahrens ein Wert von 14% berechnet, der deutlich geringer ausfällt als die Observer-Streuungen, der aber erwartungsgemäß höher liegt als bei der „In-vitro-Situation“ des Schilddrüsenphantoms (7,3%).

Die Analyse der Intraobserver-Variabilität macht deutlich, dass die individuellen Untersuchungsergebnisse – ähnlich wie bei der Phantom-Studie – eine zum Teil erhebliche Variationsbreite aufweisen, die sich zwischen -27% und +84% bewegt (Fig. 11,12). Aus dem Teildiagramm von Untersucher 3 lässt sich beispielsweise ein eindeutiger Bias der Messwerte ablesen, wobei die Mittelwerte sehr dicht am Gruppenmittelwert liegen. Die Daten von Untersucher 6 zeigen einen geringfügigen systematischen Fehler und kleine Standardabweichungen und vereinen somit eine adäquate Reproduktion der Referenzvolumina mit einer hohen Reliabilität. Bei den Messergebnissen von Untersucher 9 fallen große Differenzen der Mittelwerte (geringe Richtigkeit) und hohe Standardabweichungen (geringe Präzision) auf, ein Hinweis darauf, dass die Wiederholungsmessungen insgesamt sehr ungenau sind.
Aus Fig. 11 ist außerdem abzulesen, dass die Folgemessungen desselben Untersuchers am gleichen Probanden nicht beeinflusst werden, d.h. es ist kein Memory-Effekt erkennbar. Damit bestätigt sich, was zur Vermeidung eines Bias ohnehin im Studiendesign vorausgesetzt wurde.
Die Intraobserver-Streuung s_{intra} von 22% liegt in der Größenordnung von Angaben in der Literatur. Eine Studie von Özgen et al. [Özgen 1999] zur Schilddrüsenvolumetrie mittels Ultraschall an insgesamt 60 gesunden Kindern ergab beim gleichen Untersucher eine Variationsbreite von 22% als 95%-Limits of Agreement. Bei Lyshchik et al. [Lyshchik 2004a] lag die Intraobserver-Variabilität mit 14,4% etwas niedriger. Andere Autoren [Brunn 1981, Schlögl 2001, Reinartz 2002] stellten fest, man müsse bei der Volumetrie von großen

oder irregulär geformten Strumen mit konventionellem 2D-Ultraschall sogar individuelle Variationsbreiten von bis zu 23% - 35% akzeptieren.

Aus Gründen der Qualitätssicherung ist es interessant, die bei den Volumenmessungen durch unterschiedliche Untersucher auftretende Intraobserver-Variabilität mit deren Praxiserfahrung zu korrelieren. Dabei zeigen die vorliegenden Daten, dass sowohl die größte als auch die kleinste untersucherbezogene Variationsbreite bei als „sehr erfahren“ geltenden Anwendern auftrat. Eigentlich wäre zu erwarten gewesen, dass die Messergebnisse der erfahreneren Kollegen insgesamt weniger variieren als die der weniger geübten, was auch Jarløv et al. zeigten [Jarløv 1991]. Somit lässt dieser Befund die Forderung nahe liegend erscheinen, regelmäßige Trainingseinheiten durchzuführen, um die handwerklich-technischen Fähigkeiten aller Ultraschallanwender zu verbessern und homogenere klinische Ergebnisse zu erzielen.
Ein weiterer hilfreicher Schritt in Richtung Qualitätsverbesserung könnte neben regelmäßigen Trainingseinheiten der Einsatz des oben beschriebenen Schilddrüsenphantoms sein [Schlögl 2006]. Dieses ist derzeit für die bereits erwähnte LISA-Studie vorgesehen, mit der die medikamentöse Therapie der Knotenstruma evaluiert werden soll [Grussendorf 2005]. Bestandteil dieser Studie sind mehrmalige Scans von simulierten Schilddrüsenlappen und -läsionen sowie das entsprechende Feedback zur Scantechnik, was in den beteiligten Studienzentren ($n > 50$) den Qualitätsstandard in der Bildgebung verbessern helfen soll.

Auch bei der Interobserver-Variabilität sind die Literaturdaten konsistent mit dem eigenen Ergebnis für die Interobserver-Streuung s_{inter} von 23%. Brauer et al. [Brauer 2005] berichten in ihrer aktuellen Studie zur Schilddrüsenvolumetrie an 42 zufällig ausgewählten Erwachsenen mit asymptomatischen Schilddrüsenknoten über eine Interobserver-Variabilität von 24% für die Ellipsoidformel. Bei der Volumetrie von Schilddrüsenknoten liegt in derselben Studie der Wert noch einmal deutlich höher, nämlich bei 49%, was tendenziell bereits aus der Phantom-Studie abzuleiten war. In der oben zitierten Studie von Özgen et al. [Özgen 1999] zum Schilddrüsenvolumen bei gesunden Kindern wurde unter

Bezugnahme auf den Mittelwert von drei 2D-Ultraschallmessungen als dem „wahren“ Schilddrüsenvolumen jedes Probanden ein Interobserver-Fehler von bis zu 30% ermittelt, was sich ebenfalls in ähnlicher Dimension bewegt. Mit 15,3% liegt das Ergebnis für die Interobserver-Variabilität bei Lyshchik et al. [Lyshchik 2004a], die die Schilddrüsen von 47 Jugendlichen im 2D-Ultraschall volumetrisch untersuchten, wieder niedriger als vergleichbare Werte.

Wie also zu erwarten war, fiel die Intraobserver-Variabilität der vorliegenden Studie geringer aus als die Interobserver-Variabilität, was in Einklang steht mit den Befunden von Özgen et al. und Lyshchik et al. [Özgen 1999, Lyshchik 2004a]. Natürlich wäre es wünschenswert, wenn beim Ultraschall pro Patient immer ein und derselbe Untersucher zur Verfügung stehen würde. So ließen sich die Einflüsse der Interobserver-Variabilität vermeiden; außerdem ließe sich eine bessere Konsistenz bei der Bestimmung des Schilddrüsenvolumens erreichen. In einem Ambulanzbetrieb mit mehreren Kollegen ist diese Forderung jedoch nur schwer in die Tat umzusetzen.

Analog zur Phantom-Studie wird der von Tong et al. berechnete zufällige Untersucherfehler (Random Observer Error) mit den Literaturdaten verglichen (siehe Fig. 15). Es kann aufgezeigt werden, dass sich die Variationsbreite der eigenen Werte von 5,5% bis 18% in etwa mit dem bei Tong et al. [Tong 1998] berechneten Intervall von ca. 7% bis ca. 24% deckt. Damit ergeben sich für das 2D-Ultraschallverfahren bei Schilddrüsen und Prostatae Fehler in der gleichen Größenordnung.
Ein ähnliches Bild zeigt sich beim Intraobserver-Bias: Liegen die eigenen untersucherbezogenen Mittelwerte der relativen Differenzen zwischen 5,4% und 28%, geben Tong et al. [Tong 1998] einen Schwankungsbereich von ca. -2% bis ca. +38% an. Durch diese Daten bestätigt sich die Dimension der Fehlervariation der 2D-Ultraschallmethode.

Einen weiteren interessanten Aspekt stellen die Daten dar, die zur klinischen Evaluierung von Volumenänderungen erhoben wurden. Wie bereits ausgeführt,

ist es nämlich für klinisch-volumetrische Verlaufskontrollen essentiell zu wissen, wann eine gemessene Zu- oder Abnahme des Volumens auch als solche gelten kann und welcher Anteil bei einem gewählten Konfidenzlevel auf die Observer-Variabilität zurückzuführen ist. Für die Probanden-Schilddrüsen ließen sich auf dem 95%-Wahrscheinlichkeitsniveau minimale, sicher detektierbare Volumenänderungen von 63% (Interobserver) bzw. 61% (Intraobserver) errechnen. Bei anderen Sicherheitswahrscheinlichkeiten β ergeben sich entsprechend andere Werte für ΔV (siehe Punkt 2.4.3 in „Material und Methoden“). Bewegen sich die ΔV-Werte in der Größenordnung der Streuungen, liegt die Irrtumswahrscheinlichkeit bei 48%. Bei der Berechnung ist zu beachten, dass bei zwei Messungen immer die doppelte Varianz in die Formel eingeht. Manche Autoren [Brauer 2005] rechnen jedoch nur mit der einfachen Varianz, woraus eine zu geringe Irrtumswahrscheinlichkeit resultiert.

Wird die Diagnose einer euthyreoten diffusen Iodmangelstruma gestellt, wäre für den Kliniker zur Evaluierung einer indizierten Pharmakotherapie ein Kriterium hilfreich, anhand dessen er entscheiden kann, ob tatsächlich eine Volumenänderung vorliegt. Wie bereits erwähnt, wurde durch eine Vielzahl von Studien belegt, dass durch Präparate wie Iodid, Levothyroxin oder Kombinationen aus beiden Substanzen eine Volumenreduktion erreicht werden kann [Hintze 1985, Olbricht 1985, Pfannenstiel 1988, Hintze 1989, Einenkel 1992, Wilders-Truschnig 1993, Grussendorf 1996, Rönnefarth 1996, Förster 1998, Klemenz 1998, Kreißl 2001]. Bei den meisten dieser Untersuchungen lag die mittlere Volumenabnahme bei ca. 15% – 35%. Soll individuell, d.h. bei einem Patienten, entschieden werden, ob aufgrund zweier 2D-Ultraschallmessungen eine Reduktion in dieser Größenordnung vorliegt, geben die Formeln in Tab. 1 (siehe Punkt 2.4.3 in „Material und Methoden“) Hinweise auf die möglichen Irrtumswahrscheinlichkeiten. Bei den ermittelten relativen Intra- bzw. Interobserver-Streuungen von etwa 20% ist ein Cut-off von 40% zu empfehlen, wenn Volumenänderungen in beiden Richtungen möglich sind (zweiseitiger Test); die Irrtumswahrscheinlichkeit ist also in den meisten Fällen nicht zu vernachlässigen. Eine geringere Irrtumswahrscheinlichkeit lässt sich nur mit der Annahme erzielen, dass

nur eine Volumenreduktion möglich ist. In diesem Fall kann der Cut-off bei 20% angesetzt werden (entspricht einer Sicherheitswahrscheinlichkeit von 76% bei einseitigem Test).

Um die Ergebnisse mit denen von Tong et al. vergleichen zu können, wurden zusätzlich die Reliabilitätskoeffizienten und die Standardmessfehler (*SEM*) nach dem varianzanalytischen Modell von Eliasziw [Eliasziw 1994] bestimmt. Weil die Resultate entscheidend von der Varianz der Messobjekte abhängen, wurden die Berechnungen gleich dreimal angewandt: auf alle Probanden, auf eine "Normalgruppe" und auf die normierten Messwerte.
Für die Inter- und Intraobserver-Streuungen, die in diesem Fall als Standardmessfehler SEM_{inter} und SEM_{intra} bezeichnet werden, ergaben sich 17% bzw. 14%. Mit 22% bzw. 16% liegen die korrespondierenden Werte bei Tong et al. in einer vergleichbaren Dimension. Daraus lassen sich mit einer Sicherheitswahrscheinlichkeit β von 95% sicher detektierbare Volumenänderungen von 43% bzw. 34% errechnen (Tong et al.: 61% bzw. 43%).
Für die Reliabilitätskoeffizienten $\hat{\rho}_{intra}$ und $\hat{\rho}_{inter}$ lassen sich in der Hauptgruppe die höchsten Werte ermitteln (0,94 bzw. 0,92), da hier die Objekt-Varianz im Verhältnis zur Observer-Varianz am größten ist. Da auch bei Tong et al. – in etwa gleiche Verteilung der Objektvolumina vorausgesetzt – die Werte in der gleichen Größenordnung liegen (0,93 bzw. 0,87), bestätigt sich bei ähnlichem Studiendesign die vergleichbare Reliabilität der Methode. In der Normalgruppe ist aufgrund der größeren Homogenität die Objekt-Variabilität geringer, entsprechend deutlicher fällt die Observer-Varianz ins Gewicht und entsprechend niedriger fallen demnach auch die Reliabilitätskoeffizienten aus (0,80 bzw. 0,70). Bei Transformation der Daten auf die Referenzvolumina wird die Objekt-Varianz sehr viel kleiner, weshalb nur Werte von 0,65 und 0,50 erreicht werden.
Als ergänzender Punkt konnte durch die Varianzanalyse gezeigt werden, dass es keine signifikante Interaktion zwischen Proband und Observer gibt, das heißt z.B. dass die Untersucher keine Tendenz haben, kleine Objekte überdurchschnittlich kleiner zu messen.

Vergleicht man die direkt berechneten Streuungen s_{inter} und s_{intra} mit den Standardmessfehlern SEM_{inter} und SEM_{intra}, so fällt auf, dass letztere deutlich geringer ausfallen. Weil die entsprechenden Größen jeweils in die Formeln für ΔV_{inter} bzw. ΔV_{intra} (siehe Punkt 2.4.3 in „Material und Methoden“) eingehen, ergeben sich somit verschiedene Werte für die minimal detektierbaren Volumina, nämlich 63% und 61% bzw. 46% und 38%. Hierzu ist zu bemerken, dass statistische Kenngrößen immer von dem zugrunde liegenden Modell abhängen. Zusätzlich verlangt die Anwendung einer Varianzanalyse stärkere Voraussetzungen, die immer nur annähernd erfüllt sein können. Im Zweifel sind die konservativen Schätzungen vorzuziehen.

Radioiodtherapie

Die Bestimmung des Schilddrüsenvolumens ist auch essenzieller Bestandteil der Dosimetrie vor einer Radioiodtherapie [Peters 1995, Lucas 2000, Schlögl 2001, Reinartz 2002, Reiners 2004] und wird von manchen Autoren [Peters 1995] sogar als wichtigster Baustein für die Aktivitätsberechnungen erachtet, um eine bestimmte Dosis in Organen über 30 g zu erzielen. Aber auch im Rahmen des Therapie-Monitorings nach Radioiodtherapie besitzt die 2D-Ultraschallvolumetrie einen hohen Stellenwert, um eine Reduktion des Schilddrüsenvolumens quantitativ zu erfassen [Peters 1996]. Allerdings sollten die hier charakterisierten Variabilitäten dieses Verfahrens als Unsicherheitsfaktoren bei der Ermittlung der I-131-Therapieaktivität berücksichtigt werden. Man könnte sogar behaupten, dass Über- *und* Unterschätzungen des Schilddrüsenvolumens im 2D-Ultraschall einen wesentlichen Anteil an inadäquat berechneten therapeutischen Radioiodaktivitätsmengen haben. So führten bei Reinartz et al. [Reinartz 2002] Ultraschall-Volumenberechnungen im Vergleich mit MRT in der prätherapeutischen Dosimetrie bei 60 Patienten mit multifokaler Schilddrüsenautonomie bzw. Morbus Basedow zwar nur zu einer im Durchschnitt etwa 170 Megabecquerel (MBq) niedrigeren I-131 Aktivitätsmenge; dennoch konnten die Autoren der Studie einen negativen Effekt auf die Erfolgsrate der Radioiodtherapie nicht ausschließen, da in Einzelfällen Abweichungen von bis zu 1128 MBq beobachtet wurden.

Die hier vorgelegten Ergebnisse belegen, dass hinreichend genaue 2D-Ultraschallmessungen von der Objektgröße abhängen. Dadurch lassen sich die Resultate einer Vielzahl von Radioiodtherapien erklären, die bei unifokalen und multifokalen Schilddrüsenautonomien bessere Ergebnisse liefern als bei der disseminierten Autonomie, die meist in diffusen Strumen auftritt (bisher unveröffentlichte Daten aus der Klinik für Nuklearmedizin der Universität Würzburg). Daher ist es gerade bei großen Strumen ratsam, sich nicht allein auf die 2D-Ultraschallvolumetrie zur I-131-Aktivitätsberechnung zu verlassen. Da die größten Messfehler bei extrem großen Strumen auftreten [Hussy 2000], wäre in diesen vergleichsweise seltenen Fällen eine zusätzliche Volumetrie mittels 3D-Ultraschall, CT oder MRT hilfreich.

3D-Ultraschall

Zur Zeit werden 3D-Ultraschallscanner entwickelt, die sowohl systematische als auch zufällige Fehler bei der Ultraschallvolumetrie reduzieren und dadurch Richtigkeit und Präzision der Ultraschalldiagnostik verbessern helfen könnten. Zusätzlich können 3D-Datensätze elektronisch gespeichert, nachbearbeitet und für Verlaufskontrollen genutzt werden. Im Moment stehen einer breiten klinischen Anwendung dieser Technik jedoch noch einige größere Hürden im Weg. Dazu gehören Ungenauigkeiten durch Verwischungseffekte bei der manuellen Abgrenzung von unscharfen Organkonturen und zum Teil erhebliche Artefakte, die durch Patientenbewegung, Atem- und Schlucktätigkeit während des Scanvorgangs entstehen. Für den klinischen Routineeinsatz ist die Bildbearbeitung im 3D-Ultraschall aktuell zu zeitaufwending und arbeitsintensiv.

Die hier quantifizierte Variationsbreite des 2D-Ultraschallverfahren bestätigt die Notwendigkeit, zur Approximation der Schilddrüsenvolumens und zur Verifizierung pathogenetischer und therapeutischer Konzepte eine exakte und reproduzierbare Methode einzusetzen. Diese sollte für den Patienten nicht belastend und wenig zeitaufwendig sein. Der 3D-Ultraschall besitzt das Potenzial, die Intra- und Interobserver-Variabilität bei der 2D-Ultraschallvolumetrie zu verringern und gleichzeitig Richtigkeit und Genauigkeit der Ultraschalldiagnostik zu steigern. Trotz der Umständlichkeiten in der klinischen Routine sollte der Ein-

satz dieses Verfahrens gerade in der individuellen Dosimetrie sowie bei klinischen Studien gefordert werden. Möglicherweise könnten schon bald neue semiautomatische Bildverarbeitungstechniken die Anwendung erleichtern und dem 3D-Ultraschall zu einer besseren Akzeptanz verhelfen.

4.3 Synopsis und Ausblick

Aufgrund des unterschiedlichen Designs von Phantom- und Probanden-Studie (In-vitro- vs. In-vivo-Bedingungen) ist eine vergleichende Gegenüberstellung der Ergebnisse nur sehr eingeschränkt möglich. Hinzu kommt, dass nur Objekte gleicher Größenordnung, in diesem Fall also die Schilddrüsenlappen, verglichen werden können. Aber auch hier ist ein Vergleich insofern asymmetrisch, als beim Phantom definierte Volumina vorgegeben sind und wegen der begrenzten Anzahl an Objekten keine verlässlichen Schätzungen abgegeben werden können.
Unter Berücksichtigung dieser Limitationen können zu Richtigkeit und Präzision der Untersuchungen folgende Aussagen gemacht werden: in beiden Studien werden die Schilddrüsenlappen im 2D-Ultraschall signifikant überschätzt, wobei der Bias bei den Probanden deutlich höher liegt (17% vs. 5,0%). Wie zu erwarten war, zeigen die Inter- und Intraobserver-Streuungen bei den Probandenschilddrüsen im Vergleich zu den Phantom-Lappen eine wesentlich größere Variationsbreite (23%/22% vs. 8,3%/8,7%). Auch der Fehler der Messmethode, der in etwa gleichzusetzen ist mit dem Inner-Subjekt-Variationskoeffizienten CV, fällt bei der Phantom-Studie weniger ins Gewicht (7,3% vs. 14%). Ebenfalls etwas geringer ist die Spannweite für den rein untersucherbezogenen Messfehler (Random Observer Error) nach Tong et al. [Tong 1998]: beim Phantom von 1,6% bis 13,7%, bei den Probanden von 5,5% bis 18%. In die Parallelität dieser Befunde passt, dass bei beiden Studien der systematische Fehler unterhalb der Streuungen liegt.

In Zusammenschau von Phantom- und Probanden-Studie lässt sich schlussfolgern, dass das Volumen kleiner intrathyreoidaler Herdbefunde systematisch eher *unter*schätzt wird, bei normal großen bis gering vergrößerten Schilddrüsenlappen lässt sich sowohl im Phantom als auch bei den Probanden eine systematische *Über*schätzung feststellen.

Ist die 2D-Ultraschall-Volumetrie somit ein praxistaugliches Verfahren? Diese Frage kann aufgrund der vorliegenden Daten bejaht werden. Mit einer Interobserver-Variabilität in vivo von 23% erweist es sich als hinreichend genau, denn dieser Wert liegt im mittleren Drittel der Variationsbreite von 10% – 30%, die in den Leitlinien der Deutschen Gesellschaft für Nuklearmedizin zum Thema Schilddrüsendiagnostik als „Interobserver-Varianz“ vorgegeben ist [Dietlein 2003]. (Anmerkung: Mathematisch korrekt wäre "Interobserver-Streuung". Da der Begriff „Varianz“ als Quadrat der Streuung bereits anderweitig besetzt ist, sollte man als Maß für die Streuung richtiger von „Variabilität“ sprechen). Dies unterstreicht noch einmal die Meinung anderer Autoren, die schon früher die Reproduzierbarkeit der Ultraschalltechnik allgemein als gut bezeichneten [Rasmussen 1974, Tannahill 1978, Hansen 1979]. Schon bei der volumetrischen Bestimmung kleiner intrathyreoidaler Herdbefunde und erst recht bei großen, multinodös veränderten Strumen nehmen die Messfehler dieses Verfahrens jedoch Dimensionen an, die die Zuhilfenahme anderer Techniken, wie z.B. des 3D-Ultraschalls, hilfreich erscheinen lassen.

Aufgrund der hier beschriebenen Fehlermodalitäten ist eine gezielte Ausbildung der Ärzte auf dem Gebiet der Sonographie zu fordern, um die Intra- und Interobserver-Variabilität zu minimieren und möglichst untersucherunabhängige Ergebnisse zu erzielen. Eine Methode zum Training einer exakten sonographischen Volumetrie mit objektivierbaren Ergebnissen wäre der Einsatz des hier beschriebenen Schilddrüsenphantoms mit unterschiedlichen, genormten Volumina. Vor allem für die Hersteller von Sonographiegeräten wäre es eine wichtige Aufgabe, normierte Schilddrüsenphantome möglichst realitätsgetreu zu entwerfen und vielen Nutzern zur Verfügung zu stellen.

5. Zusammenfassung

Die 2D-Ultraschallvolumetrie der Schilddrüse und ihrer knotigen Veränderungen ist sowohl durch die Subjektivität und Erfahrung der Untersucher (Intra- und Interobserver-Variabilität), als auch durch die Messmethode selbst (reiner Messfehler) limitiert. Alle auftretenden Fehlerkomponenten bei der Bestimmung des Schilddrüsengesamtvolumens und von Herdbefunden unterschiedlicher Größe und Echogenität gehen ein in die Gesamtvarianz, deren Bestandteile (Einzelvarianzen) mit Hilfe der vorliegenden Phantom- und Probanden-Studie durch Vergleich mit Referenzwerten mathematisch umfassend charakterisiert und quantifiziert wurden.

Von folgenden Größenordnungen der Fehler ist auszugehen:

- Bei der **Phantom-Studie** werden die Volumina der simulierten Knoten statistisch signifikant unterschätzt mit einem Bias von -7,5%. Kleine Knoten liegen bei -9,6%, mittlere bei -8,0% und große bei -5,0%. Die zugehörigen mittleren Inter- bzw. Intraobserver-Streuungen s_{inter} bzw. s_{intra} liegen bei 15% bzw. 14%. Die Echogenität scheint dabei keinen wesentlichen Einfluss auf die Messgenauigkeit der Knoten zu haben. Die Schilddrüsenlappen werden dagegen von den Untersuchern um 5,0% überschätzt (ebenfalls signifikant). Der reine Fehler der Messmethode liegt bei 11% für die Knoten und bei 7,3% für die Lappen. Für die mit 95% Wahrscheinlichkeit sicher detektierbaren Volumenänderungen errechnet sich für die Knoten ein Interobserver-Wert von 41%, für die Lappen liegt er bei 23%.

- Bei der **Probanden-Studie** errechnet sich für die Schilddrüsenvolumina ein statistisch signifikanter Bias von +17%. Die zugehörigen mittleren Streuungen s_{inter} bzw. s_{intra} liegen bei 23% bzw. 22%. Interessanterweise ließ sich bei der getrennten Volumenbestimmung des rechten und linken

Schilddrüsenlappens ein signifikanter Unterschied feststellen. Für den methodenimmanenten Messfehler ergab sich in vivo ein Wert von 14%. Auf dem 95%-Wahrscheinlichkeitsniveau errechnen sich die sicher detektierbaren Volumenänderungen mit 63% für verschiedene Untersucher (Interobserver) bzw. 61% beim gleichen Untersucher (Intraobserver).
Die Ergebnisse der multivariaten Reliabilitätsanalyse können als zusätzliche Informationen zur Varianzaufklärung aufgefasst werden und sollen dazu dienen, die Zuverlässigkeit der 2D-Messungen einzuordnen.

In Übereinstimmung mit Literaturdaten kann die 2D-Ultraschallvolumetrie auf der mathematischen Basis des Ellipsoidmodells somit als praxistaugliches Verfahren gelten und erweist sich auch für große epidemiologische Studien als geeignet. Ein Vergleich mit den Leitlinien der Deutschen Gesellschaft für Nuklearmedizin lässt eine Observer-Variabilität in ähnlicher Größenordnung erwarten.

Das hier vorgestellte Schilddrüsenphantom kann behilflich sein, um die manuellen Fertigkeiten von Ärzten beim Einsatz des Ultraschalls zu trainieren, objektivierbare Ergebnisse bei der Volumetrie unterschiedlich großer Objekte zu erzielen und die Intra- und Interobserver-Variabilität zu reduzieren.
Möglicherweise erreichen auch schon bald technische Weiterentwicklungen des 3D-Ultraschallverfahrens Praxisreife, mit dessen Hilfe die Intra- und Interobserver-Variabilität bei der Schilddrüsenvolumetrie deutlich reduziert werden könnten.

Untersucher, die auf dem Gebiet der 2D-Ultraschallvolumetrie tätig sind, sollten sich der hier dargestellten Fehlermöglichkeiten auf der Basis des Ellipsoidmodells und ihrer Größenordnungen bewusst sein. Für Ultraschall-Anwender in der Schilddrüsendiagnostik stellt das hier vorgestellte statistische Auswerteverfahren ein fundiertes mathematisches Konzept dar, das Vergleiche mit Daten in der Literatur ermöglicht und dessen Ergebnisse als Referenzwerte für künftige Studien dienen können.

6. LITERATURVERZEICHNIS

1. **Billion H**. Zur Dosisberechnung bei der Radioiodtherapie der Hyperthyreose. Fortschr Geb Röntgenstr Nuklearmed 1958;88:460-464.

2. **Bland JM**. How should I calculate a within-subject coefficient of variation? http://www-users.york.ac.uk/~mb55/meas/cv.htm. 2006.

3. **Bland JM, Altman DG**. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995;346:1085-1087.

4. **Bland JM, Altman DG**. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-160.

5. **Bland JM, Altman DG**. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;8:307-310.

6. **Brander A, Viikinkoski P, Nickels J, Kivisaari L**. Thyroid gland: US screening in middle-aged women with no previous thyroid disease. Radiology 1989;173:507-510.

7. **Brander AE, Viikinkoski VP, Nickels JI, Kivisaari LM**. Importance of thyroid abnormalities detected at US screening: a 5-year follow-up. Radiology 2000;215:801-806.

8. **Brauer V, Hentschel B, Paschke R**. Euthyreote Schilddrüsenknoten: Therapieziele, Resultate und Perspektiven der medikamentösen Therapie. Dtsch Med Wochenschr 2003;128:2381-2387.

9. **Brauer VF, Eder P, Miehle K, Wiesner TD, Hasenclever H, Paschke R**. Interobserver variation for ultrasound determination of thyroid nodule volumes. Thyroid 2005;15:1169-1175.

10. **Brown MC, Spencer R**. Thyroid gland volume estimated by use of ultrasound in addition to scintigraphy. Acta Radiol Oncol Radiat Phys Biol 1978;17:337-341.

11. **Brunn J, Block U, Ruf G, Bos I, Kunze WP, Scriba PC**. Volumetrie der Schilddrüsenlappen mittels Real-time-Sonographie. Dtsch Med Wochenschr 1981;106:1338-1340.

12. **Castro MR, Caraballo PJ, Morris JC**. Effectiveness of thyroid hormone suppressive therapy in benign solitary thyroid nodules: a meta-analysis. J Clin Endocrinol Metab 2002;87:4154-4159.

13. **Celani MF, Mariani M, Mariani G**. On the usefulness of levothyroxine suppressive therapy in the medical treatment of benign solitary, solid or predominantly solid, thyroid nodules. Acta Endocrinol 1990;123:603-608.

14. **Chang FM, Hsu KF, Ko HC, Yao BL, Chang CH, Yu CH, Chen HY**. Three-dimensional ultrasound assessment of fetal liver volume in normal pregnancy: a comparison of reproducibility with two-dimensional ultrasound and a search for a volume constant. Ultrasound Med Biol 1997;23:381-389.

15. **Cheung PS, Lee JM, Boey JH**. Thyroxine suppressive therapy of benign solitary thyroid nodules: a prospective randomized study. World J Surg 1989;13:818-821.

16. **Dederichs B, Otte R, Klink JE, Schicha H**. Volumenreduktion der Schilddrüse nach Radioiodtherapie bei Patienten mit Schilddrüsenautonomie und Morbus Basedow. Nuklearmedizin 1996;35:164-169.

17. **Dietlein M, Dressler J, Grünwald F, Joseph K, Leisner B, Moser E, Reiners C, Rendl J, Schicha H, Schneider P, Schober O, Deutsche Gesellschaft für Nuklearmedizin**. Leitlinie zur Schilddrüsendiagnostik (Version 2). Nuklearmedizin 2003;42:109-115.

18. **Einenkel D, Bauch KH, Benker G**. Treatment of juvenile goitre with levothyroxine, iodide or a combination of both: the value of ultrasound grey-scale analysis. Acta Endocrinol 1992;127:301-306.

19. **Eliasziw M, Young SL, Woodbury MG, Fryday-Field K**. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther 1994;74:777-788.

20. **Fenster A, Downey DB, Cardinal HN**. Three-dimensional ultrasound imaging. Phys Med Biol 2001;46:R67-99.

21. **Fleiss JL, Shrout PE**. Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika 1978;43:259-262.

22. **Förster G, Krummenauer F, Hansen C, Beyer J, Kahaly G**. Individuell dosiertes Levothyroxin mit 150 µg Jodid versus 100 µg Levothyroxin kombiniert mit 100 µg Jodid. Dtsch Med Wochenschr 1998;123:685-689.

23. **Freitas JE, Freitas AE**. Thyroid and parathyroid imaging. Semin Nucl Med 1994;24:234-245.

24. **Fujimoto Y, Oka A, Omoto R, Hirose M**. Ultrasound scanning of the thyroid gland as a new diagnostic approach. Ultrasonics 1967;5:177-180.

25. **Gagner M, Inabnet III WB**. Endoscopic thyroidectomy for solitary thyroid nodules. Thyroid 2001;11:161-163.

26. **Gallo M, Pesenti M, Valcalvi R**. Ultrasound thyroid nodule measurements: the "gold standard" and its limitations in clinical decision making. Endocr Pract 2003;9:194-199.

27. **Gharib H, James EM, Charboneau JW, Naessens JM, Offord KP, Gorman CA**. Suppressive therapy with levothyroxine for solitary thyroid nodules. A double-blind controlled clinical study. N Engl J Med 1987;317:70-75.

28. **Gilja OH, Thune N, Matre K, Hausken T, Odegaard S, Berstad A**. In vitro evaluation of three-dimensional ultrasonography in volume estimation of abdominal organs. Ultrasound Med Biol 1994;20:157-165.

29. **Grussendorf M**. Therapie der euthyreoten Jodmangelstruma. Wirksamkeit der Kombination aus L-Thyroxin und 150 µg Jodid im Vergleich zu Mono-L-Thyroxin. Med Klin 1996;91:489-493.

30. **Grussendorf M, Vaupel R, Reiners C, Wegscheider K, LISA-Studiengruppe.** Die LISA-Studie - eine randomisierte, doppelblinde, vierarmige, plazebokontrollierte, multizentrische Studie an 1000 Patienten über die medikamentöse Therapie der Struma in Deutschland. Med Klin 2005;100:542-546.

31. **Guhlmann CA, Rendl J, Eilles C, Börner W**. The relevance of I-131 activity calculations for outcome of radioiodine therapy in functional thyroid autonomy. De Gruyter. Berlin, New York 1992.

32. **Gullu S, Gurses MA, Baskal N, Uysal AR, Kamel AN, Erdogan G**. Suppressive therapy with levothyroxine for euthyroid diffuse and nodular goiter. Endocr J 1999;46:221-226.

33. **Gutekunst R, Becker W, Hehrmann R, Olbricht T, Pfannenstiel P**. Ultraschalldiagnostik der Schilddrüse. Dtsch Med Wochenschr 1988;113:1109–1112.

34. **Hamper UM, Trapanotto V, DeJong MR, Sheth S, Caskey CI**. Three-dimensional US of the prostate: early experience. Radiology 1999;212:719-723.

35. **Hansen JM, Kampmann J, Madsen SN, Skovsted L, Solgaard S, Grytter C, Grontvedt T, Rasmussen SN**. L-thyroxine treatment of diffuse non-toxic goitre evaluated by ultrasonic determination of thyroid volume. Clin Endocrinol 1979;10:1-6.

36. **Hegedüs L**. Thyroid size determined by ultrasound. Influence of physiological factors and non-thyroidal disease. Dan Med Bull 1990;37:249-263.

37. **Hegedüs L, Bonnema SJ, Bennedbaek FN**. Management of simple nodular goiter: current status and future perspectives. Endocr Rev 2003;24:102-132.

38. **Hegedüs L, Karstrup S**. Ultrasonography in the evaluation of cold thyroid nodules. Eur J Endocrinol 1998;138:30-31.

39. **Hegedüs L, Perrild H, Poulsen LR, Andersen JR, Holm B, Schnohr P, Jensen G, Hansen JM**. The determination of thyroid volume by ultrasound and its relationship to body weight, age, and sex in normal subjects. J Clin Endocrinol Metab 1983;56:260-263.

40. **Hintze G, Emrich D, Köbberling J**. Therapy of endemic goitre: controlled study on the effect of iodine and thyroxine. Horm Metab Res 1985;17:362-365.

41. **Hintze G, Emrich D, Köbberling J**. Treatment of endemic goitre due to iodine deficiency with iodine, levothyroxine or both: results of a multicentre trial. Eur J Clin Invest 1989;19:527-534.

42. **Hussy E, Voth E, Schicha H**. Sonographische Bestimmung des Schilddrüsenvolumens - Vergleich mit operativen Daten. Nuklearmedizin 2000;39:102-107.

43. **Huysmans DA, de Haas MM, van den Broek WJ, Hermus AR, Barentsz JO, Corstens FH, Ruijs SH**. Magnetic resonance imaging for volume estimation of large multinodular goitres: a comparison with scintigraphy. Br J Radiol 1994;67:519-523.

44. **Igl W, Lukas P, Leisner B, Fink U, Seiderer M, Pickardt CR, Lissner J**. Sonographische Volumenbestimmung der Schilddrüse. Vergleich mit anderen Methoden. Nuklearmedizin 1981;20:64-71.

45. **Igl W, Seiderer M, Fink U, Lissner J**. Quantitative Volumenbestimmung der Schilddrüse mit Hilfe der Sonographie. Nucl Compact 1980;11:11-13.

46. **Iro H, Uttenweiler V, Zenk J**. Kopf-Hals-Sonographie. Springer-Verlag. Berlin Heidelberg 2000.

47. **Jarløv AE, Hegedüs L, Gjørup T, Hansen JM**. Accuracy of the clinical assessment of thyroid size. Dan Med Bull 1991;38:87-89.

48. **Jarløv AE, Nygaard B, Hegedüs L, Karstrup S, Hansen JM**. Observer variation in ultrasound assessment of the thyroid gland. Br J Radiol 1993;66:625-627.

49. **Klemenz B, Förster G, Wieler H, Kahaly G, Kaiser KP, Hansen C, Willkomm P, Ruhlmann J**. Studie zur Kombinationstherapie der endemischen Struma mit zwei unterschiedlichen Thyroxin/Iodkombinationen. Nuklearmedizin 1998;37:101-106.

50. **Knudsen N, Bols B, Bülow I, Jørgensen T, Perrild H, Ovesen L, Laurberg P**. Validation of ultrasonography for the thyroid gland for epidemiological purposes. Thyroid 1999;9:1069-1074.

51. **Knudsen N, Bülow I, Jørgensen T, Laurberg P, Ovesen L, Perrild H**. Goitre prevalence and thyroid abnormalities at ultrasonography: a comparative epidemiological study in two regions with slightly different iodine status. Clin Endocrinol 2000;53:479-485.

52. **Krasznai I, Földes J, Farkas G, Bohar L, Gönczi J**. Determination of euthyroid thyroid mass. Nucl Med Commun 1985;6:169-172.

53. **Kreißl M, Tiemann M, Hänscheid H, Rendl J, Reiners C**. Vergleich der Wirksamkeit zweier verschieden dosierter Levothyroxin-Jodid-Kombinationen in der Therapie der euthyreoten diffusen Struma. Dtsch Med Wochenschr 2001;126:227-231.

54. **Kuma K, Matsuzuka F, Yokozawa T, Miyauchi A, Sugawara M**. Fate of untreated benign thyroid nodules: results of long-term follow-up. World J Surg 1994;18:495-498.

55. **La Rosa GL, Lupo L, Giuffrida D, Gullo D, Vigneri R, Belfiore A**. Levothyroxine and potassium iodide are both effective in treating benign solitary solid cold nodules of the thyroid. Ann Intern Med 1995;122:1-8.

56. **Landis JR, Koch GG**. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

57. **Langer P**. Normal thyroid size versus goiter - postmortem thyroid weight and ultrasonographic volumetry versus physical examination. Endocrinol Exp 1989;23:67-76.

58. **Le Moli R, Wesche MF, Tiel-Van Buul MM, Wiersinga WM**. Determinants of longterm outcome of radioiodine therapy of sporadic non-toxic goitre. Clin Endocrinol 1999;50:783-789.

59. **Lima N, Knobel M, Cavaliere H, Sztejnsznajd C, Tomimori E, Medeiros-Neto G**. Levothyroxine suppressive therapy is partially effective in treating patients with benign, solid thyroid nodules and multinodular goiters. Thyroid 1997;7:691-697.

60. **Loevner LA**. Imaging of the thyroid gland. Semin Ultrasound CT MR 1996;17:539-562.

61. **Lucas KJ**. Use of thyroid ultrasound volume in calculating radioactive iodine dose in hyperthyroidism. Thyroid 2000;10:151-155.

62. **Luster M, Jacob M, Thelen MH, Michalowski U, Deutsch U, Reiners C**. Reduktion des Schilddrüsenvolumens nach Radiojodtherapie wegen funktioneller Autonomie. Nuklearmedizin 1995;34:57-60.

63. **Lyshchik A, Drozd V, Demidchik Y, Reiners C**. Diagnosis of thyroid cancer in children: value of gray-scale and power doppler US. Radiology 2005;235:604-613.

64. **Lyshchik A, Drozd V, Reiners C**. Accuracy of three-dimensional ultrasound for thyroid volume measurement in children and adolescents. Thyroid 2004a;14:113-120.

65. **Lyshchik A, Drozd V, Schlögl S, Reiners C**. Three-dimensional ultrasonography for volume measurement of thyroid nodules in children. J Ultrasound Med 2004b;23:247-254.

66. **Mainini E, Martinelli I, Morandi G, Villa S, Stefani I, Mazzi C**. Levothyroxine suppressive therapy for solitary thyroid nodule. J Endocrinol Invest 1995;18:796-799.

67. **Marinelli LD, Quinby EH, Hine GJ**. Dosage determination with radioactive isotopes. Practical considerations in therapy and protection. Am J Roentgenol 1948;59:260-281.

68. **Miccoli P, Berti P, Materazzi G, Minuto M, Barellini L**. Minimally invasive video-assisted thyroidectomy: five years of experience. J Am Coll Surg 2004;199:243-248.

69. **Miccoli P, Minuto MN, Orlandini C, Galleri D, Massi M, Berti P**. Ultrasonography estimated thyroid volume: a prospective study about its reliability. Thyroid 2006;16:37-39.

70. **Myhill J, Reeve TS, Figgins PM**. Measurement of the mass of the thyroid gland in vivo. Am J Roentgenol Radium Ther Nucl Med 1965;94:828-836.

71. **Naik KS, Bury RF**. Imaging the thyroid. Clin Radiol 1998;53:630-639.

72. **Noma S, Nishimura K, Togashi K, Itoh K, Fujisawa I, Nakano Y, Konishi J, Kasagi K, Iida Y, Itoh H, et al**. Thyroid gland: MR imaging. Radiology 1987;164:495-499.

73. **Nygaard B, Hegedüs L, Nielsen KG, Ulriksen P, Hansen JM**. Long-term effect of radioactive iodine on thyroid function and size in patients with solitary autonomously functioning toxic thyroid nodules. Clin Endocrinol 1999;50:197-202.

74. **Nygaard B, Nygaard T, Court-Payen M, Jensen LI, Søe-Jensen P, Nielsen KG, Fugl M, Hegedüs L**. Thyroid volume measured by ultrasonography and CT. Acta Radiol 2002;43:269-274.

75. **Olbricht T, Hoff HG, Benker G, Wagner R, Reinwein D**. Sonographische Volumetrie der Schilddrüse zur Verlaufskontrolle bei einer Thyroxin- und Jodidbehandlung der blanden Struma. Dtsch Med Wochenschr 1985;110:863-866.

76. **Olbricht T, Schmitka T, Mellinghoff U, Benker G, Reinwein D**. Sonographische Bestimmung von Schilddrüsenvolumina bei Schilddrüsengesunden. Dtsch Med Wochenschr 1983;108:1355-1358.

77. **Özgen A, Erol C, Kaya A, Özmen MN, Akata D, Akhan O**. Interobserver and intraobserver variations in sonographic measurement of thyroid volume in children. Eur J Endocrinol 1999;140:328-331.

78. **Papini E, Bacci V, Panunzi C, Pacella CM, Fabbrini R, Bizzarri G, Petrucci L, Giammarco V, La Medica P, Masala M, Pitaro M, Nardi F**. A prospective randomized trial of levothyroxine suppressive therapy for solitary thyroid nodules. Clin Endocrinol 1993;38:507-513.

79. **Papini E, Petrucci L, Guglielmi R, Panunzi C, Rinaldi R, Bacci V, Crescenzi A, Nardi F, Fabbrini R, Pacella CM**. Long-term changes in nodular goiter: a 5-year prospective randomized trial of levothyroxine suppressive therapy for benign cold thyroid nodules. J Clin Endocrinol Metab 1998;83:780-783.

80. **Peeters EY, Shabana WM, Verbeek PA, Osteaux MM**. Use of a curved-array transducer to reduce interobserver variation in sonographic measurement of thyroid volume in healthy adults. J Clin Ultrasound 2003;31:189-193.

81. **Peters H, Fischer C, Bogner U, Reiners C, Schleusener H**. Radioiodine therapy of Graves' hyperthyroidism: standard vs. calculated 131 iodine activity. Results from a prospective, randomized, multicentre study. Eur J Clin Invest 1995;25:186-193.

82. **Peters H, Fischer C, Bogner U, Reiners C, Schleusener H**. Reduction in thyroid volume after radioiodine therapy of Graves' hyperthyroidism: results of a prospective, randomized, multicentre study. Eur J Clin Invest 1996;26:59-63.

83. **Peters H, Fischer C, Bogner U, Reiners C, Schleusener H**. Treatment of Graves' hyperthyroidism with radioiodine: results of a prospective randomized study. Thyroid 1997;7:247-251.

84. **Pfannenstiel P**. Therapie der endemischen Struma mit Levothyroxin und Jodid. Dtsch Med Wochenschr 1988;113:326-331.

85. **Quadbeck B, Prüllage J, Roggenbuck U, Hirche H, Janssen OE, Mann K, Hörmann R**. Long-term follow-up of thyroid nodule growth. Exp Clin Endocrinol Diabetes 2002;110:348-354.

86. **Rago T, Bencivelli W, Scutari M, Di Cosmo C, Rizzo C, Berti P, Miccoli P, Pinchera A, Vitti P**. The newly developed three-dimensional (3D) and two-dimensional (2D) thyroid ultrasound are strongly correlated, but 2D overestimates thyroid volume in the presence of nodules. J Endocrinol Invest 2006;29:423-426.

87. **Rago T, Vitti P, Chiovato L, Mazzeo S, De Liperi A, Miccoli P, Viacava P, Bogazzi F, Martino E, Pinchera A**. Role of conventional ultrasonography and color flow-doppler sonography in predicting malignancy in 'cold' thyroid nodules. Eur J Endocrinol 1998;138:41-46.

88. **Rasmussen SN, Hjorth L**. Determination of thyroid volume by ultrasonic scanning. J Clin Ultrasound 1974;2:143-147.

89. **Reinartz P, Sabri O, Zimny M, Nowak B, Cremerius U, Setani K, Büll U**. Thyroid volume measurement in patients prior to radioiodine therapy: comparison between three-dimensional magnetic resonance imaging and ultrasonography. Thyroid 2002;12:713-717.

90. **Reiners C**. Functional autonomy of the thyroid: volume reduction after radioiodine treatment. Exp Clin Endocrinol Diabetes 1993;101:136-138.

91. **Reiners C, Becker W, Hagemann J, Köhler F**. Schilddrüsendiagnostik - Nutzen bildgebender Verfahren. Teil 1: Diagnostische Aussage. Fortschr Med 1987a;105:125-128.

92. **Reiners C, Becker W, Hagemann J, Köhler F**. Schilddrüsendiagnostik - Nutzen bildgebender Verfahren. Teil 2: Initial- und Verlaufsdiagnostik. Fortschr Med 1987b;105:163-166.

93. **Reiners C, Schneider P**. Radioiodine therapy of thyroid autonomy. Eur J Nucl Med Mol Imaging 2002;29 Suppl 2:S471-478.

94. **Reiners C, Wegscheider K, Schicha H, Theissen P, Vaupel R, Wrbitzky R, Schumm-Dräger PM**. Prevalence of thyroid disorders in the working population of Germany: ultrasonography screening in 96,278 unselected employees. Thyroid 2004;14:926-932.

95. **Reinhardt MJ, Brink I, Joe AY, von Mallek D, Ezziddin S, Palmedo H, Krause TM**. Radioiodine therapy in Graves' disease based on tissue-absorbed dose calculations: effect of pre-treatment thyroid volume on clinical outcome. Eur J Nucl Med Mol Imaging 2002a;29:1118-1124.

96. **Reinhardt MJ, Joe A, von Mallek D, Zimmerlin M, Manka-Waluch A, Palmedo H, Krause TM**. Dose selection for radioiodine therapy of borderline hyperthyroid patients with multifocal and disseminated autonomy on the basis of Tc-99m-pertechnetate thyroid uptake. Eur J Nucl Med Mol Imaging 2002b;29:480-485.

97. **Reverter JL, Lucas A, Salinas I, Audi L, Foz M, Sanmarti A**. Suppressive therapy with levothyroxine for solitary thyroid nodules. Clin Endocrinol 1992;36:25-28.

98. **Riccabona M, Nelson TR, Pretorius DH, Davidson TE**. Distance and volume measurement using three-dimensional ultrasonography. J Ultrasound Med 1995;14:881-886.

99. **Riccabona M, Nelson TR, Pretorius DH, Davidson TE**. In vivo three-dimensional sonographic measurement of organ volume: validation in the urinary bladder. J Ultrasound Med 1996;15:627-632.

100. **Richter B, Neises G, Clar C**. Pharmacotherapy for thyroid nodules. A systematic review and meta-analysis. Endocrinol Metab Clin North Am 2002;31:699-722.

101. **Rönnefarth G, Kauf E, Deschner F, Forberger M**. Therapie der Jodmangelstruma bei Jugendlichen mit Jodid oder einer Kombination von Jodid und Levothyroxin unter besonderer Berücksichtigung der Lipidparameter. Klin Pädiatr 1996;208:123-128.

102. **Schlögl S, Andermann P, Luster M, Reiners C, Lassmann M**. A novel thyroid phantom for ultrasound volumetry - a determination of intra- and inter-observer variability. Thyroid 2006;16:41-44.

103. **Schlögl S, Werner E, Lassmann M, Terekhova J, Muffert S, Seybold S, Reiners C**. The use of three-dimensional ultrasound for thyroid volumetry. Thyroid 2001;11:569-574.

104. **Schmitz W**. Radiojodtherapie der Hyperthyreosen unter besonderer Berücksichtigung der Dosisberechnung. Ärztl Forsch 1963;17:237-240.

105. **Schumm PM, Strohm WD, Maul FD, Kirchner C, Usadel KH**. Schilddrüsen-Volumenbestimmung. Inn Med 1982;9:166-169.

106. **Shabana W, Peeters E, De Maeseneer M**. Measuring thyroid gland volume: should we change the correction factor? AJR Am J Roentgenol 2006;186:234-236.

107. **Shabana W, Peeters E, Verbeek P, Osteaux MM**. Reducing interobserver variation in thyroid volume calculation using a new formula and technique. Eur J Ultrasound 2003;16:207-210.

108. **Solbiati L, Croce F**. Thyroid and parathyroid glands. Abdominal and General Ultrasound 1993:661-676.

109. **Stuart A, Ord JK**. Kendall's advanced theory of statistics. Hodder Arnold. London 1994.

110. **Szebeni A, Beleznay E**. New simple method for thyroid volume determination by ultrasonography. J Clin Ultrasound 1992;20:329-337.

111. **Tan GH, Gharib H**. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. Ann Intern Med 1997;126:226-231.

112. **Tannahill AJ, Hooper MJ, England M, Ferriss JB, Wilson GM**. Measurement of thyroid size by ultrasound, palpation and scintiscan. Clin Endocrinol 1978;8:483-486.

113. **Tong S, Cardinal HN, McLoughlin RF, Downey DB, Fenster A**. Intra- and inter-observer variability and reliability of prostate volume measurement via two-dimensional and three-dimensional ultrasound imaging. Ultrasound Med Biol 1998;24:673-681.

114. **Van Isselt JW, de Klerk JM, van Rijk PP, van Gils AP, Polman LJ, Kamphuis C, Meijer R, Beekman FJ**. Comparison of methods for thyroid volume estimation in patients with Graves' disease. Eur J Nucl Med Mol Imaging 2003;30:525-531.

115. **Vitti P, Martino E, Aghini-Lombardi F, Rago T, Antonangeli L, Maccherini D, Nanni P, Loviselli A, Balestrieri A, Araneo G, et al**. Thyroid volume measurement by ultrasound in children as a tool for the assessment of mild iodine deficiency. J Clin Endocrinol Metab 1994;79:600-603.

116. **Wemeau JL, Caron P, Schvartz C, Schlienger JL, Orgiazzi J, Cousty C, Vlaeminck-Guillem V**. Effects of thyroid-stimulating hormone suppression with levothyroxine in reducing the volume of solitary thyroid nodules and improving extranodular nonpalpable changes: a randomized, double-blind, placebo-controlled trial by the French Thyroid Research Group. J Clin Endocrinol Metab 2002;87:4928-4934.

117. **Wesche MF, Tiel-van Buul MM, Smits NJ, Wiersinga WM**. Ultrasonographic versus scintigraphic measurement of thyroid volume in patients referred for 131I therapy. Nucl Med Commun 1998;19:341-346.

118. **Wiedemann W, Reiners C**. Die Differentialdiagnose des echoarmen Knotens der Schilddrüse. Dtsch Med Wochenschr 1982;51:1972-1975.

119. **Wienke JR, Chong WK, Fielding JR, Zou KH, Mittelstaedt CA**. Sonographic features of benign thyroid nodules: interobserver reliability and overlap with malignancy. J Ultrasound Med 2003;22:1027-1031.

120. **Wilders-Truschnig MM, Warnkross H, Leb G, Langsteger W, Eber O, Tiran A, Dobnig H, Passath A, Lanzer G, Drexhage HA**. The effect of treatment with levothyroxine or iodine on thyroid size and thyroid growth stimulating immunoglobulins in endemic goitre patients. Clin Endocrinol 1993;39:281-286.

121. **Wirtz M, Caspar F**. Beurteilerübereinstimmung und Beurteilerreliabilität. Hogrefe-Verlag. Göttingen 2002.

122. **Woodward B, Warwick R**. How safe is diagnostic sonar? Br J Radiol 1970;43:719-725.

123. **Zelmanovitz F, Genro S, Gross JL**. Suppressive therapy with levothyroxine for solitary thyroid nodules: a double-blind controlled clinical study and cumulative meta-analyses. J Clin Endocrinol Metab 1998;83:3881-3885.

7. PUBLIKATIONEN AUS DER DISSERTATION

- **Andermann P, Schlögl S, Mäder U, Luster M, Lassmann M, Reiners C.** Intra- and interobserver variability of thyroid volume measurements in healthy adults by 2D versus 3D ultrasound. Nuklearmedizin 2007;46:1-7.

- **Schlögl S, Andermann P, Luster M, Reiners C, Lassmann M.** A novel thyroid phantom for ultrasound volumetry: determination of intraobserver and interobserver variability. Thyroid 2006;16:41-46.

8. ABKÜRZUNGSVERZEICHNIS

2D / 3D	zweidimensional / dreidimensional
ANOVA	Analysis of Variance
B-Mode	Brightness-Mode
cm	Zentimeter
CT	Computertomographie
CV	Coefficient of Variation
D	Differenz
dB	Dezibel
f	Koeffizient
GHz	Gigahertz
gr	groß
Gy	Herddosis in Gray
I-131	Iod-131
ICC	Intra Class Correlation
J	Joule
kHz	Kilohertz
kl	klein
Kn	Knoten
m	männlich
ml	Milliliter
MHz	Megahertz
MRT	Kernspintomographie
MS	Mean Squares (mittlere Quadratsummen)
mtl	mittelgroß
MVA	multiplanare Volumenapproximation
mm	Millimeter
MW	Mittelwert
p	Irrtumswahrscheinlichkeit
ROI	Region of Interest
S-VHS	Super Video Home System
s	Streuung
Std-Fehler	Standardfehler
SEM	Standard Error of Measurement
US	Ultraschall
V	Volumen
VAR	Varianz
w	weiblich

DANKSAGUNG

Mein aufrichtiger Dank gilt Herrn Prof. Dr. Chr. Reiners, der mir das Thema der Arbeit zur Verfügung gestellt hat und mir durch seine kritische, wohlwollende Begleitung und seine Motivation ein äußerst wertvoller Mentor war.

Herrn Priv.-Doz. Dr. M. Beer danke ich ganz herzlich für die Übernahme des Korreferates.

Ganz besonders möchte ich Herrn Dr. U. Mäder für die Betreuung der Arbeit danken, für seine Anregungen, seinen Ideenreichtum und seinen mathematischen Scharfsinn. Ohne ihn wäre ich nie in die höheren Sphären der Statistik vorgestoßen.

Herzlichen Dank an Frau Dipl. Phys. Susanne Schlögl, ihre Expertise und Unterstützung beim 3D-Ultraschall und beim Handling des Schilddrüsenphantoms.

Dank auch an alle Kolleginnen und Kollegen der Klinik für Nuklearmedizin der Universität Würzburg, die sich bereitwillig als Versuchskaninchen zur Verfügung gestellt haben.

Nicht zuletzt danke ich vor allem meiner großartigen Familie und allen lieben Freunden und Kollegen für ihre moralische Unterstützung. Sie alle waren mir ein unverzichtbarer Rückhalt und haben so zum Gelingen der Arbeit beigetragen.

Lebenslauf

Persönliche Daten

Vor- und Zuname:	Paul Andermann
Geburtsdatum:	06.07.1970
Geburtsort:	Amberg
Staatsangehörigkeit:	deutsch
Familienstand:	ledig
Eltern:	Paul Andermann
	Christine Andermann, geb. Stiewe

Schulbildung

09/1976 - 07/1980	Grundschule Kümmersbruck
09/1980 - 06/1989	Erasmus-von-Rotterdam-Gymnasium Amberg
24.06.1989	Allgemeine Hochschulreife (Note: 1,0)

Wehrdienst

06/1989 - 08/1990	Grundwehrdienst: Sanitätsbataillon 8 in Regensburg, dann Heeresmusikkorps 4 in Regensburg

Hochschulbildung

1990/91	Universität Würzburg (Anglistik, Philosophie)
05/1991	Studienbeginn Humanmedizin (Universität Würzburg)
03/1992	Aufnahme in die Studienstiftung des deutschen Volkes
04/1992	Aufnahme in das Cusanuswerk
03/1994 - 07/1994	Medizinstudium an der Universität Wien
08/1994	Erster Abschnitt der Ärztlichen Prüfung
03/1999	Zweiter Abschnitt der Ärztlichen Prüfung
06/2000	Dritter Abschnitt der Ärztlichen Prüfung
14.06.2000	Erlaubnis für die Tätigkeit als Arzt im Praktikum
01.08.2002	Approbation als Arzt

Praktisches Jahr

Chirurgie:	Chirurgische Universitätsklinik Würzburg
Innere Medizin:	Medizinische Universitätsklinik Würzburg
	Royal North Shore Hospital, Sydney, Australien
Dermatologie:	Groote Schuur Hospital, Kapstadt, Südafrika
	Mount Sinai School of Medicine, New York, USA

Berufsausbildung

02/2001 – 07/2002	Arzt im Praktikum (AiP) an der Nuklearmedizinischen Klinik am Klinikum rechts der Isar der TU München
08/2002 – 12/2006	Assistenzarzt an der Klinik und Poliklinik für Nuklear-medizin der Universität Würzburg
27.07.2006	Facharzt für Nuklearmedizin